IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 24, 2019, accepted December 10, 2019, date of publication December 16, 2019,
date of current version December 27, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2960037

Bioinformatics Methodologies to Identify Interactions Between Type 2 Diabetes and Neurological Comorbidities

MD HABIBUR RAHMAN[1,2,3], SILONG PENG[1,2], XIYUAN HU[1,2], CHEN CHEN[1,2], SHAHADAT UDDIN[4], JULIAN M. W. QUINN[5], AND MOHAMMAD ALI MONI[5,6]

[1]Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100190, China
[3]Department of Computer Science and Engineering, Islamic University, Kushtia 7003, Bangladesh
[4]Complex Systems Research Group and Project Management Program, Faculty of Engineering, The University of Sydney, Sydney, NSW 2008, Australia
[5]Bone Biology Division, Garvan Institute of Medical Research, Darlinghurst, NSW 2010, Australia
[6]School of Medical Sciences, Faculty of Medicine and Health, The University of Sydney, Sydney, NSW 2008, Australia

Corresponding author: Mohammad Ali Moni (mohammad.moni@sydney.edu.au)

This work was supported in part by the National Natural Science Foundation of China under Grant 61571438, and in part by the Chinese Academy of Sciences (CAS)-The World Academy of Sciences (TWAS) President's Fellowship under Grant 2016CTF014.

ABSTRACT Type 2 diabetes (T2D) is a chronic metabolic disorder characterised by high blood sugar and insulin insensitivity which greatly increases the risk of developing neurological diseases (NDs). The co-existence of T2D and comorbidities such as NDs can complicate or even cause the failure of standard treatments for those diseases. Comorbidities can be both causally linked and influence each other's development through genetic, molecular, environmental or lifestyle-based risk factors that they share. For T2D and NDs, such underlying common molecular mechanisms remain elusive but large amounts of molecular data accumulated on these diseases enable analytical approaches to identify their interconnected pathways. Here, we propose a framework to explore possible comorbidity interactions between a pair of diseases using a bioinformatic examination of the cellular pathways involved and explore this framework for T2D and NDs with analyses of a large number of publicly available gene expression datasets from tissues affected by these diseases. We designed a bioinformatics pipeline to analyse, utilize and combine gene expression, Gene Ontology (GO) and molecular pathway data by incorporating Gene Set Enrichment Analysis and Semantic Similarity. Our bioinformatics methodology was implemented in R, available at https://github.com/HabibUCAS/T2D_Comorbidity. We identified genes with altered expression shared by T2D and NDs as well as GOs and molecular pathways these diseases share. We also computed the proximity between T2D and neurological pathologies using these genes and GO term semantic similarity. Thus, our method has generated new insights into disease mechanisms important for both T2D and NDs that may mediate their interaction. Our bioinformatics pipeline could be applied to other co-morbidities to identify possible interactions and causal relationships between them.

INDEX TERMS Bioinformatics, comorbidities, gene set enrichment analysis, gene ontology, neurological disease, pathway, semantic similarity, Type 2 diabetes.

I. INTRODUCTION

Type 2 diabetes (T2D) is a complex, chronic disorder whose causation is correspondingly multifactorial and heterogeneous. Many people developing T2D go through a stage of obesity-associated insulin resistance prior to the development of frank hyperglycemia [1]. The pancreatic islets respond to insulin resistance by increasing their cell mass and insulin secretion but if this fails to compensate, chronic high levels of blood sugar (i.e., frank T2D) persists. High blood sugar levels result in glycation products that cause vascular inflammation and blockages, resulting in long-term complications and organ damage [1], [3], [24]. There are a number of hypothesized mechanisms (that are not mutually exclusive) to explain how insulin resistance and islet beta-cell dysfunction and T2D occurs and how it affects tissues. Glucotoxicity due

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

to hyperglycaemia may impair insulin secretion [4], [5] and cause beta-cell death or dysfunction [6]–[8]. Lipotoxicity due to increased serum triglycerides may impair β-cell secretory function [9], [10] or cause apoptosis [11], [12]. Oxidative stress may lead to the generation of reactive oxygen species to which beta cells are particularly vulnerable [13]–[15]. Endoplasmic reticulum (ER) stress may result from high insulin production causing the elevated beta-cell flux of proteins through the ER [16]–[18], and this may drive insulin resistance [19]. Amyloid deposition in islets may also occur which can also reduce insulin sensitivity [20]. The relative contribution of these mechanisms is often unclear, but it is evident that a range of cell pathways involving many genes are involved in T2D development.

Genetic research has identified some genes associated with the risk of developing T2D. Genetic linkage based studies have identified two genes of particular importance, namely CAPN10 and TCF7L2. Candidate gene-based studies revealed numerous other genes that include PPARG, IRS1, IRS-2, KCNJ11, WFS1, HNF1A, HNF1B, HNF4A, RAPGEF1 and TP53. Moreover, genome-wide association studies show 38 genes associated with T2D [21]. One study identified ten loci (MTNR1B, SLC30A8, THADA, TCF7L2, KCNQ1, CAMK1D, CDKAL1, IGF2BP2, HNF1B, and CENTD2) to be associated with reduced beta-cell function, and three loci (PPARG, FTO and KLF14) associated with reduced insulin sensitivity [22].

Thus, the intrinsic multifactorial nature of T2D marks it a very complex disease to understand and treat effectively [3], [23]–[27]. However, an important clinical issue is that there are many diseases that commonly occur in association with T2D, notably NDs [28]–[36]; as comorbidities of T2D, ND development may be caused or exacerbated by the activation of molecular pathways and disease-causing genes they share with T2D. This includes pathways and genes involved in inflammation and response to high glycation products or those involved in responses to glucose and lipid toxicity, oxidative and ER stress and amyloid deposition. The co-existence of two or more such serious diseases can greatly complicate treatments for individual diseases and, indeed, greatly elevate mortality [37].

As much multi-omics data has been released into the public domain, analysing such data to identify common pathways shared by different diseases that can occur as comorbidities have become an important use of bioinformatics. Nevertheless, the available data and information are usually deliberated in isolation and rarely combined due to the lack of an appropriate bioinformatics approach. This lack of combination studies results in overlooking important disease-causing factors that act synergistically to cause the complex comorbidity. The integration of different bioinformatics tools can provide methodologies that extract more discriminating information from the available data on disease interactions.

As T2D results in an increased risk of NDs, it is likely that these have shared molecular factors that enable their interaction in the same individual. Since these underlying common molecular mechanisms remain elusive, in this study we designed and implemented a bioinformatic pipeline to utilize and combine gene expression, gene ontology (GO) and molecular pathway data by incorporating Gene Set Enrichment Analysis and Semantic Similarity studies. In addition, we have incorporated semantic similarity approaches to use genes and GO terms to measure the proximity of the comorbidities to identify biological processes important to each disease. Moreover, we performed the verification of the results with gold benchmark databases and literature.

II. METHODOLOGY

A. SURVEY OF AVAILABLE DATA

We retrieved gene expression microarray data for this study from the public repository available at National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/geo/) [38]. Queries for T2D initially returned 1547 datasets from which most of the datasets were discarded considering the following issues and finally selected datasets with a minimum number of 6 samples to exploit the maximum power of the study. We have employed the following criteria for selecting the datasets for our study:

1) Studies in which samples from an earlier study were re-analysed or investigated with the different processes were not considered for our study.
2) Some datasets are related to T2D or NDs, but samples used were of a type (e.g., type of cell extract) irrelevant for our study were not considered.
3) Only data from human subjects were used.
4) We considered only those datasets whose differentially expressed genes (DEGs) count is greater than 50 when applying a threshold value of absolute 1 for log fold change (logFC).

Gene expression datasets for T2D disease yielded the following datasets from GEO repositories: GSE20966, a study of 10 control and 10 T2D subjects from pancreatic tissue [39]; GSE23343, a study of 7 control subject and 10 T2D subjects using liver biopsies [40]; GSE25724, a study of 7 control and 6 T2D subjects using pancreatic islet tissue [41]; GSE29221, an analysis of 3 control and 3 T2D subjects using skeletal muscle cell samples [42]; GSE29226, a study of 3 control and 3 T2D subjects using subcutaneous tissue samples [43]; GSE29231, a study of 3 control and 3 T2D subjects using visceral adipose tissue [44]; and GSE55650, a study of control and T2D subjects from muscle cell samples. It should be noted that for the last dataset, the data was divided into two parts based on extraction and separate analysis of myoblasts and myotubes; for myoblasts, there were 6 control and 6 T2D subjects and for myotubes, 5 control and 6 T2D [45].

For the comorbidity interactions analysis, our selected studies on neurological diseases used the following datasets. For Alzheimer Disease (AD) we used dataset GSE1297, a study of 9 control and 22 AD subjects using hippocampal CA1 tissue [46]; GSE4226 and GSE4229, which included

studies of peripheral blood cells using normal elderly control (NEC) and AD subjects [47], [48]; GSE12685, a study of 8 control and 6 AD subjects using frontal cortex synaptoneurosome samples [49]; and GSE28146, a study of 8 control and 22 AD subjects using laser captured CA1 tissue [50]. To study Parkinson's Disease (PD) we used dataset GSE7621, a study of 9 control and 16 PD subjects from substantia nigra tissue from both male and female subjects [51]; GSE19587, a study of 5 control and 6 PD subjects using post-mortem brain tissue samples [52]; GSE20141, an analysis of 8 control and 10 PD subjects also using the post-mortem brain tissue samples [53]; GSE20333, an analysis of 6 control and 6 PD subjects using substantia nigra tissue samples; GSE22491, an analysis of 8 control and 10 PD subjects using peripheral blood samples [54]; GSE28894, a study of tissue samples from four different brain regions with control and PD subjects; GSE42966, analysis of 6 control and 9 PD subjects using substantia nigra tissue samples; and GSE54536, a study of 5 control and 5 PD subjects using peripheral blood samples [55], [56]. For the case of Amyotrophic Lateral Sclerosis (ALS) we used dataset GSE833, a study of 4 control and 7 sporadic and familial ALS subjects using post mortem spinal cord [57]; GSE4595, an analysis of 9 control and 11 sporadic ALS subjects using human motor cortex tissue samples [58]; GSE19332, a study of 7 control and 3 sporadic ALS subjects using motor neuron tissue samples [59]; GSE52672, an analysis of 10 control and 10 sporadic ALS and familial ALS subjects using whole spinal cord homogenate [60]; and GSE68605, a study of 3 control and 8 ALS subjects using motor neuron tissue samples [61]. To study Epileptic Diseases (ED) we used dataset GSE22779, a study of 4 control and 12 ED subjects using mononuclear blood cells [62]; and GSE32534, a study of 5 control and 5 ED subjects using peritumoral neocortex tissue samples [63]. To study Huntington's Disease (HD) we used dataset GSE1751, a study of 14 control and 17 HD subjects using human blood samples [64]; GSE1767, a study of 14 control and 17 HD subjects using lymphocyte cells extracted from blood samples [65]; GSE24250, a study of whole blood samples using 6 control and 8 HD subjects [66]; and GSE77558, an analysis of 6 control and 6 HD subjects from striatum-like tissue samples [67]. To study Cerebral Palsy (CP) we used dataset GSE11686, a study of 4 control and 12 CP subjects using muscle biopsies [68]; and GSE31243, a study of 20 control and 20 CP subjects using skeletal muscle biopsies [69]. To study Multiple Sclerosis (MS) we used dataset GSE7102, a study of 6 control and 6 treated MS subjects; GSE16461, a study of 8 control and 8 MS subjects using CD4+ T cells from blood samples [70]; GSE17048, a study of 54 control and 99 MS subjects using peripheral blood cell samples [71]; GSE21942, a study of 15 control and 12 MS subjects using peripheral blood samples [72]; GSE26484, a study of 4 control and 3 MS subjects using peripheral blood samples [73]; GSE32915, a study of 4 control and 12 MS subjects using brain tissue [74]; GSE37750, a study of 8 control, 9 before treatment and 9 after treatment subjects using extracted plasmacytoid dendritic cells (pDC) [75]; GSE52139, a study of 8 control and 8 MS subjects using spinal cord samples [76]; and GSE103005, a study of 12 control and 8 MS subjects using whole blood samples.

B. GENE SET ENRICHMENT ANALYSIS

Gene set enrichment analysis (GSEA) is an analytical method based on statistical approaches that interpret gene expression data to identify a set of DEGs with altered expression levels [77]. These genes may be interconnected with disease phenotypes. GSEA uses a set of previously grouped genes having common biological pathway involvement or chromosomal location. It compares genes obtained from two cell categories through DNA microarray or next-generation sequencing (NGS) by analysing their expression levels in different conditions or disease states. The gene set that falls in the extremes of this list: up-regulated and down-regulated are considered to be associated with the phenotypic differences.

1) PATHWAYS

Molecular pathway comprised of a series of actions within the molecules of the human cells that cause a certain product or change in the cell. This kind of pathway triggers the assembly of new molecules. Moreover, a pathway can also turn genes on or off. To make insights into the molecular pathways of T2D that overlap with NDs, we employed KEGG databases [78] to identify molecular pathways enriched by the DEGs.

2) ONTOLOGIES

Gene ontology (GO) is a conceptual model of gene product functions that can give important information about those systems or pathways that are significantly involved in a disease or biological function. GO datasets represent an ongoing project aims to give and ever-more comprehensive and updated structured information about biological systems [79]. GO has three domains: cellular component, molecular functions and biological process (BP). We focused on BP in this study.

3) SEMANTIC SIMILARITY

To measure the proximity of the genes and GO terms, we incorporated the semantic similarity. Semantic similarity is a method to measure proximity using ontologies to estimate the closeness between terms/concepts by defining a topological similarity [80]. This is a graph-based approach using directed acyclic graphs (DAGs) of terms (genes, GO). The semantics of these terms depends on its position in the DAG and its semantic contribution factor with all of its ancestor terms.

Formally, a GO term P can be represented as a graph $DAG_P = (P, T_P, E_P)$, where T_P is the set of all GO terms in DAG_P as ancestor terms of P together with term P itself in the GO graph and E_P is the set of corresponding edges that connect the GO terms in DAG_P. The semantic value of GO

term P is numerically calculated as,

$$\begin{cases} S_P(P) = 1 & t = P \\ S_P(t) = \max\left\{w_e * S_T(t') | t' \in \text{children of}(t)\right\} & t \neq T \end{cases} \quad (1)$$

where w_e is the semantic contribution multiplier for edge e ($e \in T_P$), generic term t with its child term t'. The semantic contribution is assigned between 0 and 1 according to the type of association. The global semantic value for P is calculated as

$$SV(P) = \sum_{t \epsilon T_p} S_p(t) \quad (2)$$

Now, if $DAG_P = (P, T_P, E_P)$ and $DAG_Q = (Q, T_Q, E_Q)$ are two terms P and Q respectively then the semantic similarity between P and Q is

$$sim(P, Q) = \frac{\sum_{t \in T_p \cap T_Q} (S_p(t) + S_Q(t))}{SV(P) + SV(Q)} \quad (3)$$

Given two sets of terms $P_1 = \left\{t_{11,t_{12},\ldots t_{1k}}\right\}$ and $Q_1 = \left\{t_{21,t_{22},\ldots t_{2n}}\right\}$, where the first set of terms length is k and the second set of terms length is n. To calculate the semantic similarity, we applied the best-match average (BMA) [81] for two given sets as follows:

$$sim_{BMA}(P_1, Q_1) = \frac{\sum_{i=1}^{k} \max_{1 \leq j \leq n} \left\{t_{1i}, t_{2j}\right\} + \sum_{j=1}^{n} \max_{1 \leq i \leq k} \left\{t_{1i}, t_{2j}\right\}}{k + n} \ldots \quad (4)$$

with i, j indices on P, Q terms.

C. THE PIPELINE OF THE PROPOSED METHODOLOGY

Figure 1 shows the steps of the pipeline of the proposed methodology:

- **DATA RETRIEVAL**. At first, we downloaded the selected GEO datasets along with their associated platform and phenotype information and transformed this into an expression set class object. We considered disease affected patients and healthy controls for T2D for identifying DEGs;
- **MODEL DESIGN**. We reviewed the GSM records manually, classified samples and created design models. The design model was T2D vs control for T2D and ND vs control or ND vs treated for the NDs were the designed model;
- **LINEAR and BAYESIAN MODEL for DIFFERENTIAL EXPRESSION**. A linear and a bayesian method [82] are applied for filtering the design model using three parameters p-value, adjusted p-value (False Discovery Rate) and logFC. Using a threshold of at most 0.05 and at least 1.0 absolute value for p-value and logFC respectively, differentially expressed genes are identified;
- **GO TERMS TEST**. In this stage, we build the class topGOdata by using selected significant genes and stipulating the GO domain along with specifying the annotation

Algorithm Pseodocode for Algorithm

Input: Microarray datasets of 'S' samples containing three different conditions; case samples, control samples and treated samples. So, notionally we have X (dataset with S samples) $= X_{case} \cup X_{control} \cup X_{treated}$.

Output: Differentially expressed gene set, GO terms tree, semantic similarity matrix, Kegg enrichment graph.

1) For each dataset $i = 1, 2, \ldots, N$:
   a) Load dataset
   b) Design matrix model
      – Convert GSE datasets into expression set class
      – Create design matrix: case vs control or case vs treated
   c) Fit Linear and Bayesian model for filtering the design model.
   d) Calculate differentially expressed genes
      – Assign P-value and logFC
      – Apply False discovery rate (FDR)
      – List significant genes
      – Create and save statistical table
   e) Gene Ontology annotation
      – Create topGO class with annotation
      – Perform Fisher's exact test
      – Create and save GO terms tree
      – Create and save correspondence genes-GO terms
2) Calculate Semantic similarity
   – Load correspondence genes-GO terms files
   – List genes and GO terms
   – Select GO field and prepare annotation
   – Compute semantic similarity:
   • **for** i = 1 to k **do**
   • **for** j = 1 to n **do**
   • $\alpha = \sum_{1=1}^{k} \max_{1 \leq j \leq n} \left\{t_{1i}, t_{2j}\right\}$
   • $\beta = \sum_{j=1}^{n} \max_{1 \leq i \leq k} \left\{t_{1i}, t_{2j}\right\}$
   • **end for**
   • **end for**
   • **return** $(\alpha + \beta)/(k + n)$
   – Create genes and GO semantic similarity matrix
   – plot and save semantic similarity matrix
3) Cluster comparison
   – Load GENE list
   – Load KEGG pathway datasets
   – Select enrichKEGG function
   – Create Kegg Enrichment graph
   – plot and save Kegg Enrichment graph
4) Results
   – Differentially expressed genes list
   – GO terms tree
   – Semantic similarity matrix
   – Kegg enrichment graph

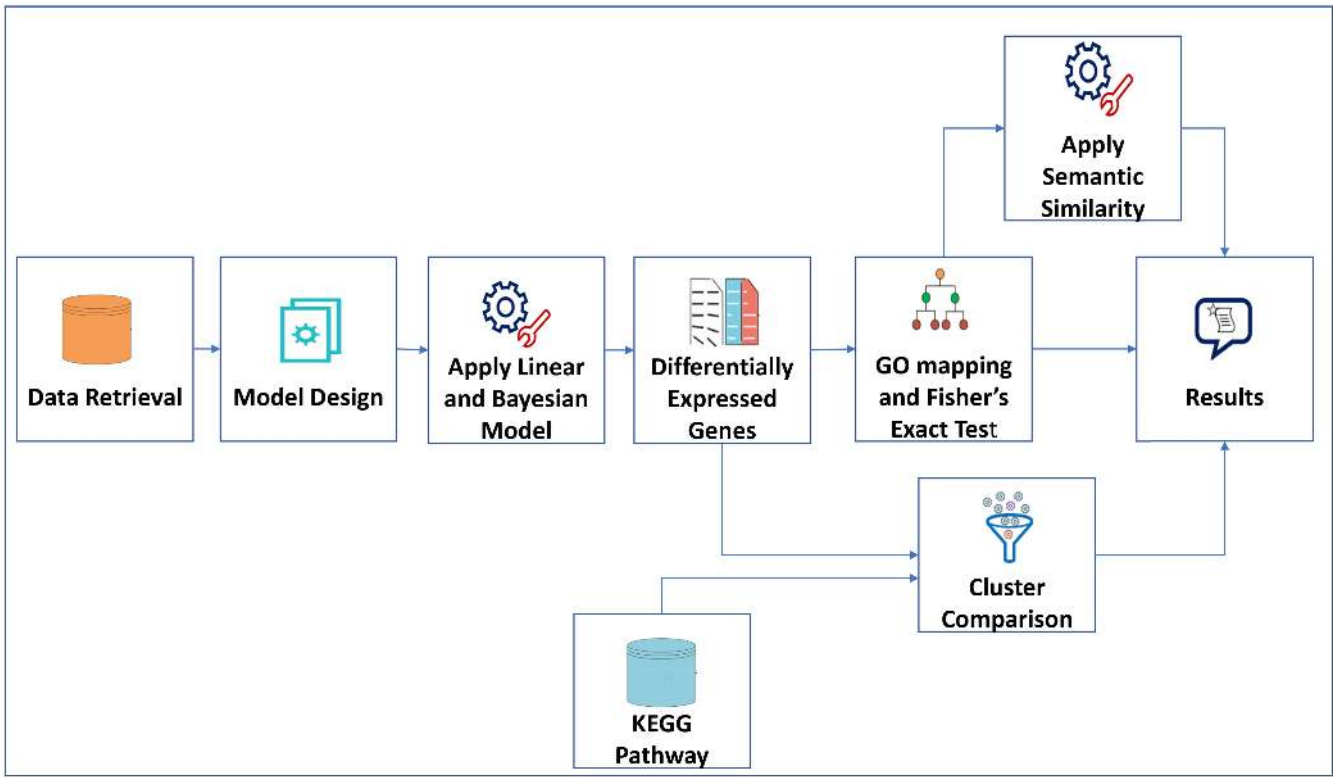


FIGURE 1. Block diagram of the proposed bioinformatics methodology.

for mapping. To obtain filtered GO terms and their relationships with genes, we employed Fisher's exact test [83].

- **SEMANTIC SIMILARITY**. After carried out the mapping, we performed the semantic similarity comparisons by means of genes and GO terms to measure the proximity among all selected T2D and ND datasets;
- **CLUSTER COMPARISON**. In this stage, we also carry out a KEGG pathway [78] enrichment test for the differentially expressed genes to identify a significant molecular pathway for T2D and its neurological comorbidities;
- **RESULTS**. The output of the pipeline of bioinformatics methodology consists of statistical summary, correspondences of genes-GO terms, DAG, gene semantic similarity matrix (and dendrogram), GO semantic similarity matrix (and dendrogram), and KEGG enrichment pathways/pathologies list. In addition, we have constructed a gene network using significant DEGs that are common to T2D and NDs along with the most prevalent pathway related to the selected pathologies.

The proposed integrated pipeline of the bioinformatics methodology is implemented in R language available at the link: https://github.com/HabibUCAS/T2D_Comorbidity. Following Bioconductor packages [84], we developed the proposed methodology using GEOquery [85] for downloading GEO datasets and expression set class transformation; limma [82] for differentially expressed gene identification from microarray data; genefilter [86] for filtering genes; topGO [83] for building the topology of DAG and identifying the significant GO terms; GOSemSim [87] for measuring the proximity among selected pathologies; clusterProfiler [88] for the KEGG pathways enrichment analysis. Eventually, the version of the used software and packages are: R 3.5.1, R Studio 1.0.136, Bioconductor 3.8, GEOquery 2.50.5, limma 3.38.3, genefilter 1.64.0, topGO 2.34.0, GOSemSim 2.8.0, DOSE 3.8.2, clusterProfiler 3.10.1.

III. RESULTS

A. STATISTICS AND GENE COMPARISON

Since each dataset has two/three conditions such as control vs case or control vs treated and after performing the statistical test for these conditions applying limma [82], we get differentially expressed genes. The statistical summary for all the selected T2D studies is tabulated in Table 1.

Table 2 summarizes common neurological disease comorbidities for T2D. The dataset legend: Alzheimer's Disease (AD): GSE1297, GSE4226, GSE4229, GSE12685, GSE28146; Amyotrophic Lateral Sclerosis (ALS): GSE833, GSE4595, GSE19332, GSE52672, GSE68605; Cerebral Palsy (CP): GSE11686, GSE31243; Epilepsy Disease (ED): GSE22779, GSE32534; Huntington Disease (HD): GSE1751, GSE1767, GSE24250, GSE77558; Multiple Sclerosis (MS): GSE7102, GSE16461, GSE17048, GSE21942,

TABLE 1. Statistical summary result for T2D disease.

Dataset	Cell source	Case Samples	Control Samples	Raw genes	P-value	logFC	Raw GSEA	Fisher GSEA
GSE20966	Pancreas	10	10	61295	7299	455	6552	4883
GSE23343	Liver Biopsy	10	7	54613	4198	962	6530	5953
GSE25724	Pancreatic islets	6	7	22283	10315	1603	6552	4883
GSE29221	Skeletal Muscle	3	3	48803	6796	2645	6571	6522
GSE29226	Subcutaneous Adipose	3	3	48803	6900	1412	6576	6423
GSE29231	Viseral Adipose	3	3	48803	9038	2078	6575	6522
GSE55650a	Skeletal Muscle cell	6	6	54613	3719	245	6530	4110
GSE55650b	Skeletal Muscle cell	6	5	54613	3032	369	6530	4765

TABLE 2. Type 2 diabetes-related neurological disease summary with the parameter value of 0.05 for p-value and an absolute value of 1 for logFC along with raw GO terms and the filtered GO terms.

Dataset	Cell source	Case Samples	Control Samples	Raw genes	P-value	logFC	Raw GSEA	Fisher GSEA
GSE1297	Hippocampal CA1 tissue	22	9	22283	2313	238	6094	3998
GSE4226	Peripheral Blood	14	14	9600	507	105	4254	3810
GSE4229	Peripheral Blood	18	22	9600	360	50	4254	3571
GSE12685	Frontal Cortex	6	8	22283	3251	149	6094	3244
GSE28146	Hippocampal CA1 tissue	22	8	54675	3405	1107	6530	6206
GSE833	Spinal cord gray matter	7	4	7070	733	489	4897	4895
GSE4595	Motor Cortex	11	9	41675	632	431	6544	5855
GSE19332	Cervical Spinal cord	3	7	54675	4564	2644	6530	6528
GSE52672	Spinal cord homogenate	10	10	526	163	80	4643	1464
GSE68605	motor neurons	8	3	54675	3171	1973	6530	6492
GSE11686	Muscle Biopsy	12	4	22383	3517	881	6094	5754
GSE31243	Skeletal muscle Biopsy	20	20	22277	6676	509	6094	4768
GSE22779	Peripheral blood	12	4	54675	15613	1514	6849	5933
GSE32534	Pertitumoral neocortex tissue	5	5	54675	2048	451	6530	5174
GSE1751	Blood Expression	17	14	22283	6217	1193	6094	5567
GSE1767	Human Blood	17	14	19881	10373	1742	6242	5908
GSE24250	Venous cellular whole blood	8	6	22283	740	402	6094	5475
GSE77558	iPSCs derived GABA MS-like neurons	6	6	47310	4155	293	6687	4170
GSE7102	Jurkat T-cells	6	6	22277	1327	443	6094	5560
GSE16461	CD4+ T cells	8	8	54675	2349	1106	6468	6424
GSE17048	Blood cell mRNA transcript (warn)	99	54	37804	3092	160	6515	3832
GSE21942	Peripheral blood mononuclear cells	12	15	54675	15415	468	6530	4436
GSE26484	Peripheral blood	3	4	54675	4448	1290	6530	6409
GSE32915	White matter brain tissue	12	4	45015	517	202	6560	6497
GSE37750a	Plasmacytoid dentric cells (pDC)	9	8	54675	6008	197	6530	3494
GSE37750b	Plasmacytoid dentric cells (pDC)	8	8	54675	5111	182	6530	3077
GSE52139	Spinal Cord (periplaque regions)	8	8	37217	2703	1501	6405	6327
GSE103005	Whole blood	8	12	47220	2157	724	6525	6309
GSE7621	Postmortem human brain	16	9	54339	5949	885	6530	5712
GSE20141	laser-dissected SNpc neurons	10	8	54675	7094	3614	6530	6530
GSE20333	Post-mortem human brains	6	6	8793	684	406	5570	5164
GSE22491	Peripheral blood mononuclear cells	10	8	41000	15015	1144	6581	5801
GSE42966	Post-mortem substantia nigra	9	6	45015	950	512	6560	6323
GSE19587	Post-mortem Brain	6	5	22277	4150	2453	6094	6071
GSE28894	Parkinson's disease brain	55	59	22184	3943	265	6510	4710
GSE54536	Peripheral Blood	5	5	47231	4522	801	6200	6071

GSE26484, GSE32915, GSE37750a, GSE37750b, GSE52139, GSE103005; Parkinson Disease (PD): GSE7621, GSE20141, GSE20333, GSE22491, GSE42966, GSE19587, GSE28894 and GSE54536.

In Table 1 and Table 2, P-values are the results of applying a statistical test that takes into account the mean difference and the variance and also the sample size. So, choosing a cut off of 0.05 means there is a 5% chance that we make the wrong decision. It implies that 5% of the tests found to be statistically significant that is false positives is accepted in this study. The sixth column values of Table 1 and Table 2 indicate the number of differentially expressed genes with the cut-off p-value of 0.05. In the GSEA, the number of statistically significant genes based on the cut-off p-value of 0.05 is found

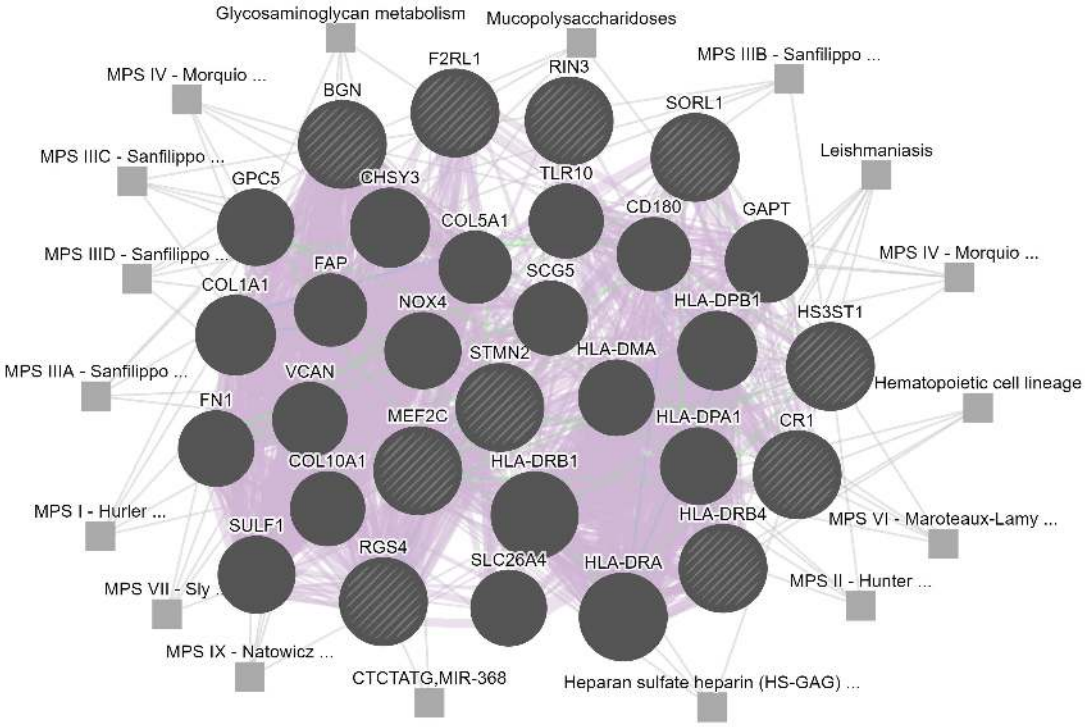
TABLE 3. The synopsis of DEGs from the different organ of the body through the proposed methodology.

Disease	Total Disease Data set	Organ /Tissue	Selected Dataset	DEGs UP	DEGs DOWN
T2D	1574	Pancreas	GSE20966, GSE25724	36	61
T2D	1547	Muscle	GSE29221, GSE55650a, GSE55650b	47	78
T2D	1547	Adipose	GSE29226, GSE29231	33	42
T2D	1547	Liver	GSE23343	32	18
AD	2501	Brain	GSE1297, GSE28146, GSE12685	93	50
AD	2501	Blood	GSE4226, GSE4229	0	3
ALS	358	Brain	GSE833, GSE4595, GSE19332, GSE52672, GSE68605	147	95
CP	484	Muscle	GSE11686,GSE31243	54	43
ED	1032	Blood	GSE22779	32	16
ED	1032	Brain	GSE32534	10	31
HD	42	Blood	GSE1751, GSE1767, GSE24250	133	12
HD	42	Brain	GSE77558	34	13
MS	3354	Blood	GSE17048, GSE21942, GSE26484, GSE103005	101	70
MS	3354	T cell	GSE7102, GSE16461	65	32
MS	3354	Lymphoid	GSE37750a, GSE37750b	76	17
MS	3354	Brain	GSE32915, GSE52139	23	50
PD	1954	Brain	GSE7621, GSE20141, GSE20333, GSE42966, GSE19587, GSE28894	175	83
PD	1954	Blood	GSE22491, GSE54536	29	56

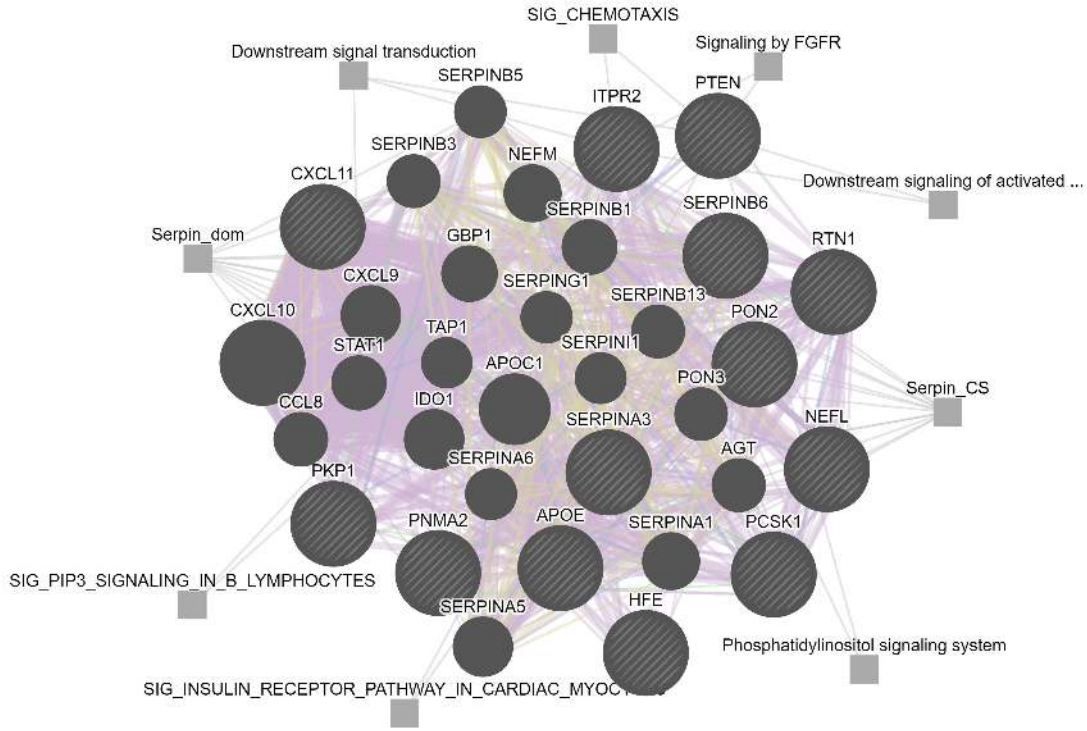
with up and down-regulated expression levels. Log2 fold change (logFC) is the log-ratio of a gene's expression levels in two different conditions such as control vs case or control vs treated and it is used to measure the changes in expression level. In the seventh column of Table 1 and Table 2, we tabulated the number of differentially expressed genes with the cut-off logFC value of 1. For GO enrichment analysis, finding raw GSEA is the first stage of the GO enrichment analysis. In this stage, we performed gene to gene ontology mapping using differentially expressed genes from the biological process (BP) domain of the gene ontology. The eight column values in Table 1 and Table 2 indicates the number of annotated gene ontology (GO) terms from the differentially expressed genes. After mapping the gene to GO terms, we applied 'Fisher's exact test' [83] statistics to get statistically significant GO terms. Fisher's exact test is based on gene counts. We performed a classical enrichment analysis with Fisher's exact test by testing the over-representation of GO terms within the group of differentially expressed genes. The ninth column is the Fisher GSEA and its values in Table 1 and Table 2 indicate the number of significant GO terms enriched from the Fisher exact test.

Table 3 narrates the summarised results of all the datasets along with the number of analysed DEGs considering the top 50 genes for each dataset.

After identifying top the 50 DEGs for all datasets, we compared the T2D and each ND. The comparison of the differential expressed genes between the T2D data sets and the neurological diseases shows that the following genes are held in common: RIN3, CR1, SORL1, MEF2C, HS3ST1, RGS4, HLA-DRB4, STMN2, F2RL1, BGN, PON2, APOE, HFE, ITPR2, SERPINA3, PTEN, SERPINB6, PCSK1, PNMA2, RTN1, NEFL, XCL11, PKP1, TENM1, CCL2, CHL1, GNAS, GREM1, MYH1, IGFBP5, TNC, GATA6, AKT3, UBE3A, GATM, SYN1, CNTNAP2, LYZ, COL6A3, MGP, UCHL1, MATR3, C9orf72, TRPM7, CHMP2B, SS18L1, TBK1, SLC25A36, HLA-DRB1, CD24, IL2RA, HLA-DQB1, IL4, CSMD1, RAB3IP, MAFB, HLA-DRB5, S100A12, NFIB, CXCL8, IFI44L, EBF2, PLA2G6, VPS13C, ATP6AP2, SCARB2, TMEM163, DGKQ, MAOA, SREBF1, SYT11, RIT2, GSTZ1, SNCA, BST1. A cluster network with this list of common genes is constructed using the online tool GeneMania [89] considering co-expression, consolidated pathways, co-localization, shared protein domain, predicted and physical interaction as shown in Figure 2. Figure 2 comprised of seven networks with common DEGs between (a) T2D and AD, (b) T2D and ALS, (c) T2D and CP, (d) T2D and ED, (e) T2D and HD, (f) T2D and MS and (g) T2D and PD. The most prevalent pathways related to the selected pathologies and their percentile coverage found in the seven networks are: mucopolysaccharidoses (0.77%), heparan sulfate heparin (HS-GAG) (2.14%), hematopoietic cell lineage (1.39%), glycosaminoglycan metabolism (0.77%), SIG PIP3 SIGNALING IN B LYMPHOCYTES (2.96%), phosphatidylinositol signaling system (1.28%), signaling by FGFR (0.62%), SIG CHEMOTAXIS (2.22%), downstream signaling of activated FGFR2 (0.68%), downstream signal transduction (0.61%), SIG INSULIN RECEPTOR PATHWAY IN CARDIAC MYOCYTES (2.05%), BIOCARTA CCR5 PATHWAY (4.55%), cell adhesion molecules (CAMs) (0.89%), tuberculosis (0.64%), influenza A (0.66%), adaptive immune system (0.57%), signaling by PDGF (0.71%), costimulation by the CD28 (2.01%), innate immune system (0.28%), BIOCARTA TH1TH2 PATHWAY (5.91%), asthma (5.29%), hematopoietic cell lineage (4.70%), allograft rejection (3.70%), translocation of ZAP-70 to immunological synapse (3.25%), intestinal immune network for IgA production (3.08%), phosphorylation of CD3 and TCR zeta chains (2.76%), autoimmune thyroid disease (2.63%), leishmaniasis (2.13%), tyrosine metabolism (1.73%), PD-1 signaling (1.39%), toxoplasmosis (1.30%), generation of

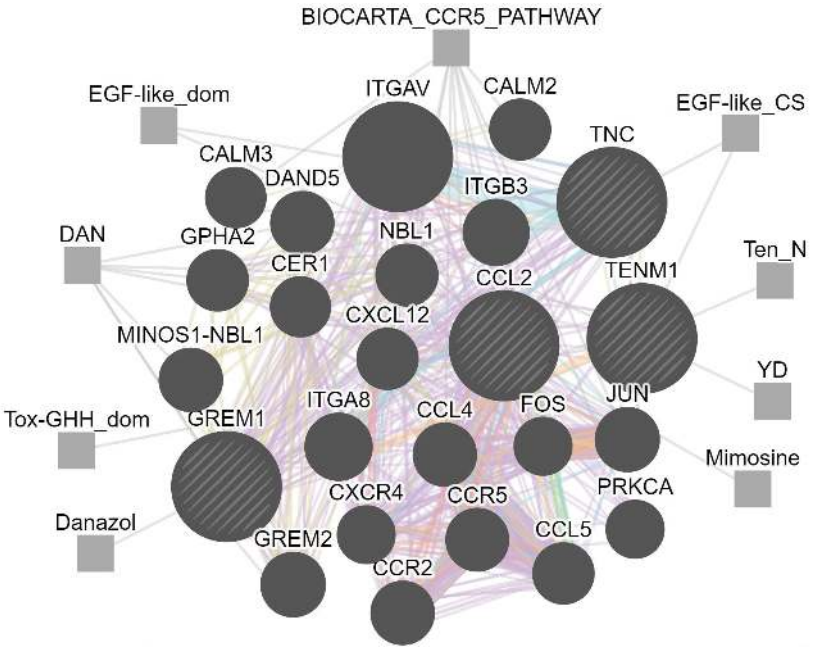


(a) Network on common DEGs between T2D and AD

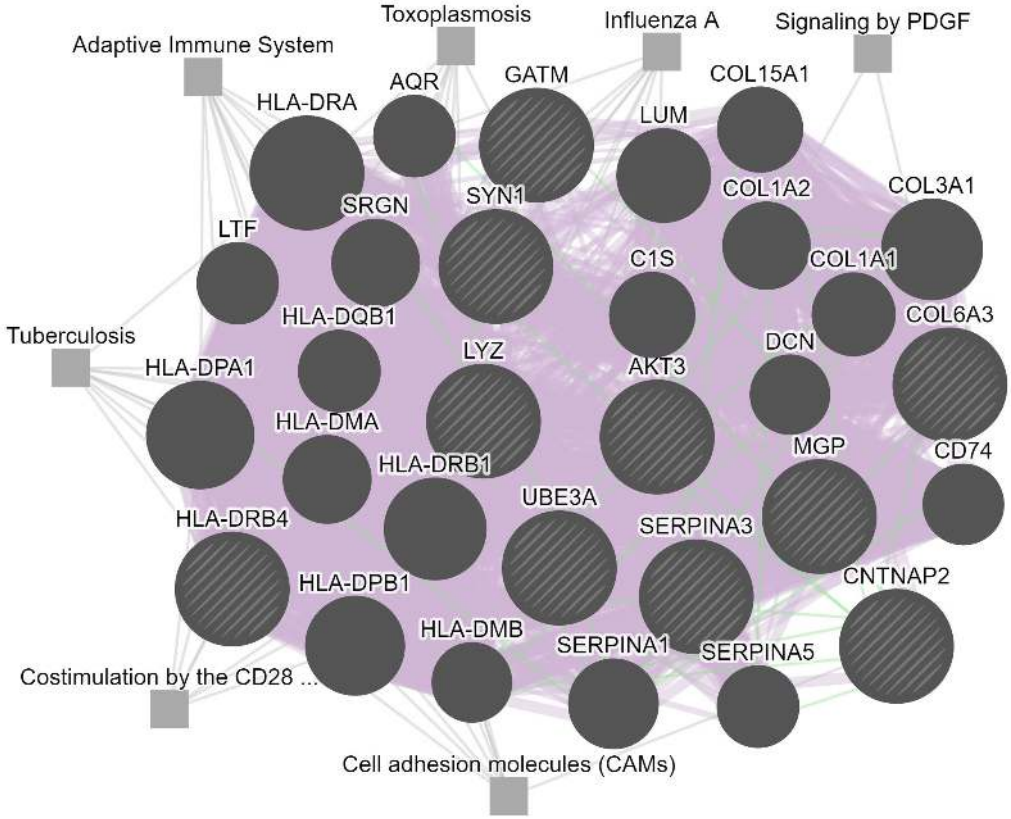


(b) Network on common DEGs between T2D and ALS

FIGURE 2. **Cluster networks with common DEGs between (a) T2D and AD (b) T2D and ALS (c) T2D and CP (d) T2D and ED (e) T2D and HD (f) T2D and MS (g) T2D and PD. The colour legends comprised of gray for consolidated pathways, pink for physical interactions, light violet for co-expression, dark yellow for predicted interactions, indigo for co-localization, light green for genetic interaction, beige for shared protein domain and light blue for the other pathways. The circle legends are striated circles for common genes and normal circle for genes added after enrichment.**

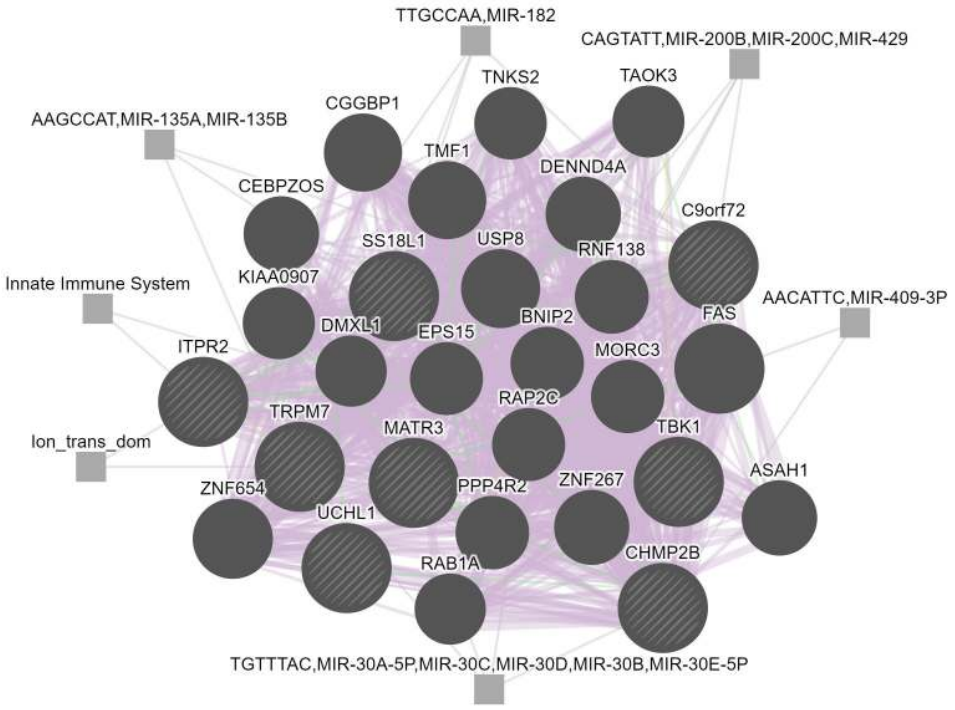


(c) Network on common DEGs between T2D and CP

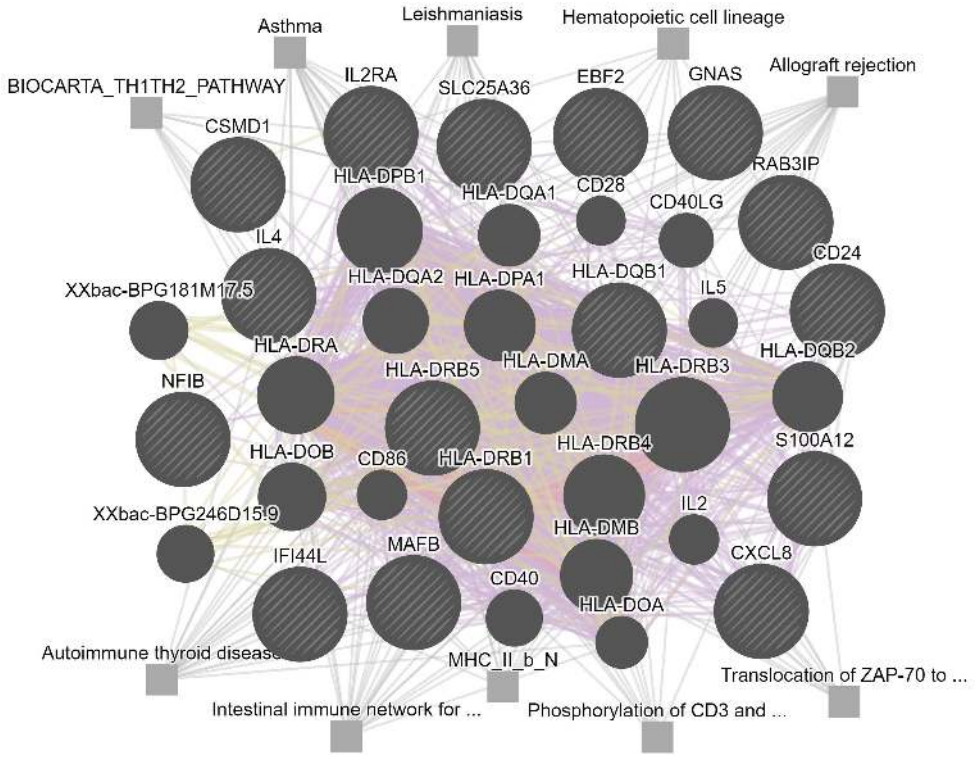


(d) Network on common DEGs between T2D and ED

FIGURE 2. *(Continued.)* **Cluster networks with common DEGs between (a) T2D and AD (b) T2D and ALS (c) T2D and CP (d) T2D and ED (e) T2D and HD (f) T2D and MS (g) T2D and PD. The colour legends comprised of gray for consolidated pathways, pink for physical interactions, light violet for co-expression, dark yellow for predicted interactions, indigo for co-localization, light green for genetic interaction, beige for shared protein domain and light blue for the other pathways. The circle legends are striated circles for common genes and normal circle for genes added after enrichment.**

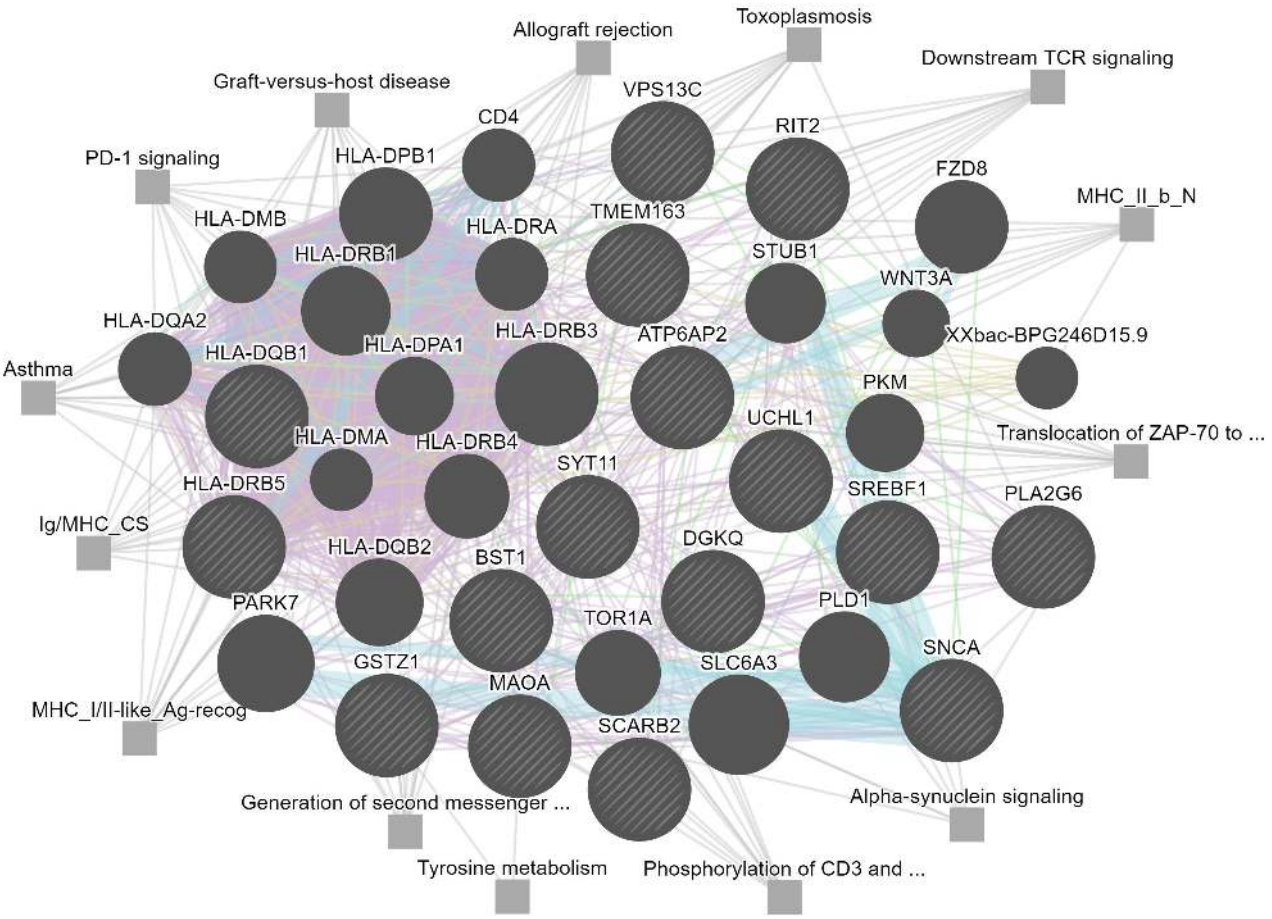


(e) Network on common DEGs between T2D and HD



(f) Network on common DEGs between T2D and MS

FIGURE 2. ***(Continued.)*** **Cluster networks with common DEGs between (a) T2D and AD (b) T2D and ALS (c) T2D and CP (d) T2D and ED (e) T2D and HD (f) T2D and MS (g) T2D and PD. The colour legends comprised of gray for consolidated pathways, pink for physical interactions, light violet for co-expression, dark yellow for predicted interactions, indigo for co-localization, light green for genetic interaction, beige for shared protein domain and light blue for the other pathways. The circle legends are striated circles for common genes and normal circle for genes added after enrichment.**



(g) Network on common DEGs between T2D and PD

FIGURE 2. *(Continued.)* **Cluster networks with common DEGs between (a) T2D and AD (b) T2D and ALS (c) T2D and CP (d) T2D and ED (e) T2D and HD (f) T2D and MS (g) T2D and PD. The colour legends comprised of gray for consolidated pathways, pink for physical interactions, light violet for co-expression, dark yellow for predicted interactions, indigo for co-localization, light green for genetic interaction, beige for shared protein domain and light blue for the other pathways. The circle legends are striated circles for common genes and normal circle for genes added after enrichment.**

second messenger molecules (0.86%), alpha-synuclein signaling (0.65%), downstream TCR signaling (0.62%), graft-versus-host disease (0.61%).

B. PATHWAY ENRICHMENT

The pathway-based analysis is a recently developed approach to understand how complex diseases may be related to each other through their underlying molecular mechanisms. After identifying DEGs, we performed KEGG pathway enrichment analysis. The relationships between KEGG pathways and all the selected datasets are represented in Figure 3. Resulting pathways in common between T2D and neurological pathologies with at least two evidence include: Influenza A, NOD-like receptor signaling pathway, Cytokine-cytokine receptor interaction, Chagas disease (American trypanosomiasis), Toll-like receptor signaling pathway, Human cytomegalovirus infection, Malaria, Autoimmune thyroid disease, Tuberculosis, PI3K-Akt signaling pathway, Melanoma, Human papillomavirus infection, Neuroactive ligand-receptor interaction, Epstein-Barr virus infection, Taste transduction, Focal adhesion, ECM-receptor interaction, Protein digestion and absorption, Relaxin signaling pathway, EGFR tyrosine kinase inhibitor resistance, cAMP signaling pathway and Amoebiasis.

C. GO ENRICHMENT AND GO TERMS TREE

The suggested biological process involved in each dataset using DEGs on type 2 diabetes disease is as follows:

- GSE20966: humoral immune response, humoral immune response mediated by circulating immunoglobin, regulation of complement activation, complement activation, cellular response to zinc ion, neuron projection extension involved in neuron projection guidance and central nervous system neuron development;
- GSE23343: myofibril assembly, cell communication by electrical coupling, astrocyte development, chemokine secretion, cell aggregation, and G protein-coupled receptor internalization;
- GSE25724: regulation of the cellular amino acid metabolic process, posttranscriptional regulation of gene expression, anaphase-promoting complex-dependent

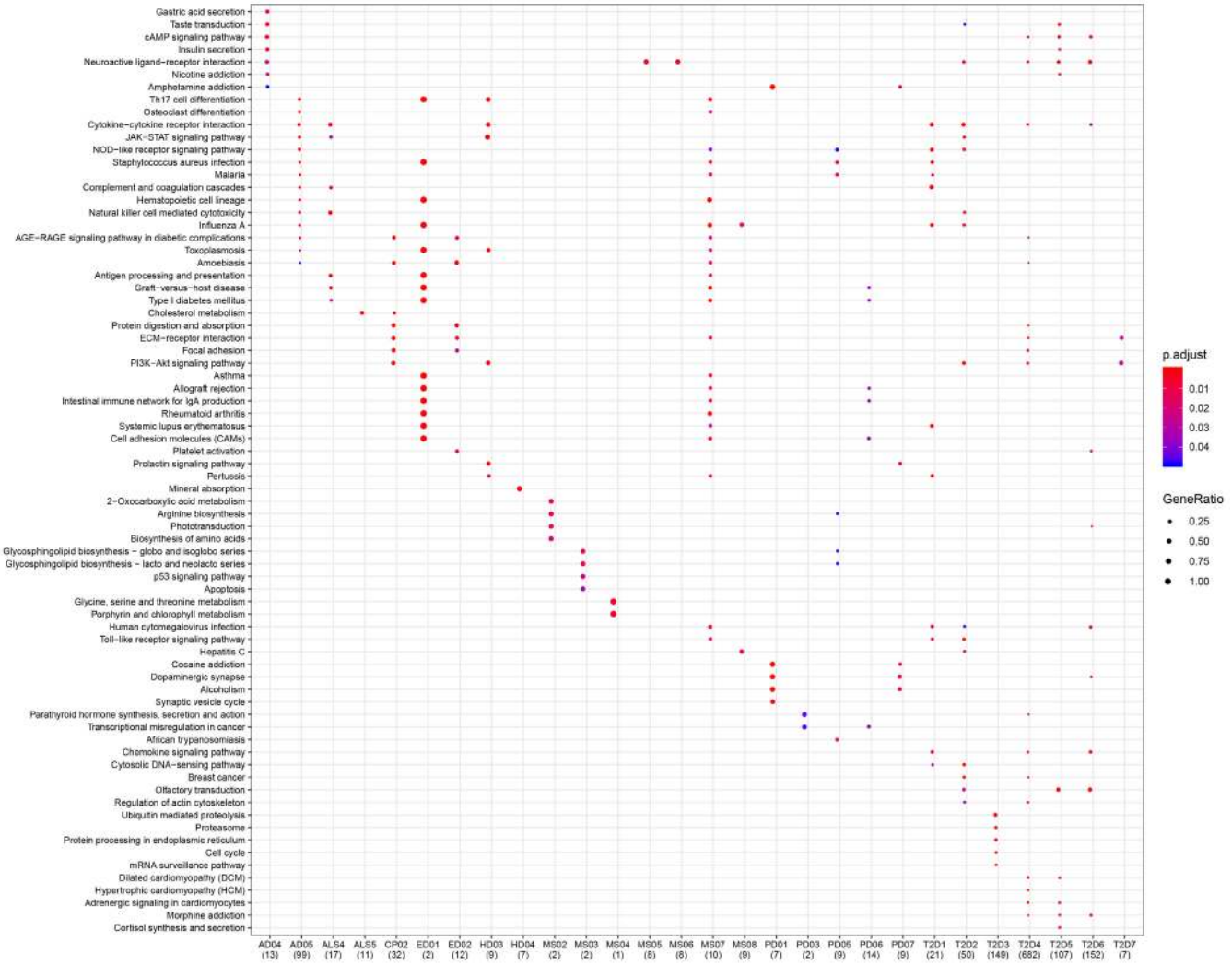


FIGURE 3. KEGG pathway enrichment analysis for differentially expressed genes. Each row represents a KEGG pathway associated with the diseases shown in columns. The domination of genes in the pathway indicated by the dimension of the circles and the range of the circles represents the statistical validation for p-value of 0.05.

catabolic process, regulation of mRNA catabolic process and cellular protein catabolic process;

- GSE29221: extracellular structure organization, extracellular matrix organization, blood circulation, collagen fibril organization, muscle system process and G protein-coupled receptor signaling pathway;
- GSE29226: sensory perception of chemical stimulus, neurological system process involved in the regulation of systemic arterial blood pressure, regulation of T-helper 2 cell cytokine production, nervous system process, and complement activation, lectin pathway;
- GSE29231: trophoblast giant cell differentiation, regulation of metanephros development, developmental induction, ERBB2 signaling pathway, regulation of ERBB signaling pathway and cell proliferation;
- GSE55650a: regulation of cell migration, nuclear division, regulation of cell motility, cell migration, cell motility, and biological adhesion;
- GSE55650b: circulatory system development, extracellular matrix organization, striated muscle cell differentiation, extracellular structure organization, anatomical structure morphogenesis and cellular response to type I interferon.

Direct Acyclic Graphs (DAG) for each selected pathology were constructed using a classic algorithm [83]. Figure 4 shows how the significant GO terms are distributed over the GO graph hierarchy with the 5 most significant GO terms (GO: 007087: cellular response to chemical stimulus, GO: 0010033: response to organic substance, GO: 0071310: cellular response to organic substance, GO: 0009653: anatomical structure morphogenesis and GO: 0071294: cellular response to zinc ion). Since the test statistic can return a p-value, we can refer them as criteria to select statistically significant GO terms. After mapping genes to GO terms, we applied Fisher's exact test statistics and classic algorithms to deal with GO enrichment analysis and GO graph structure. Fisher's

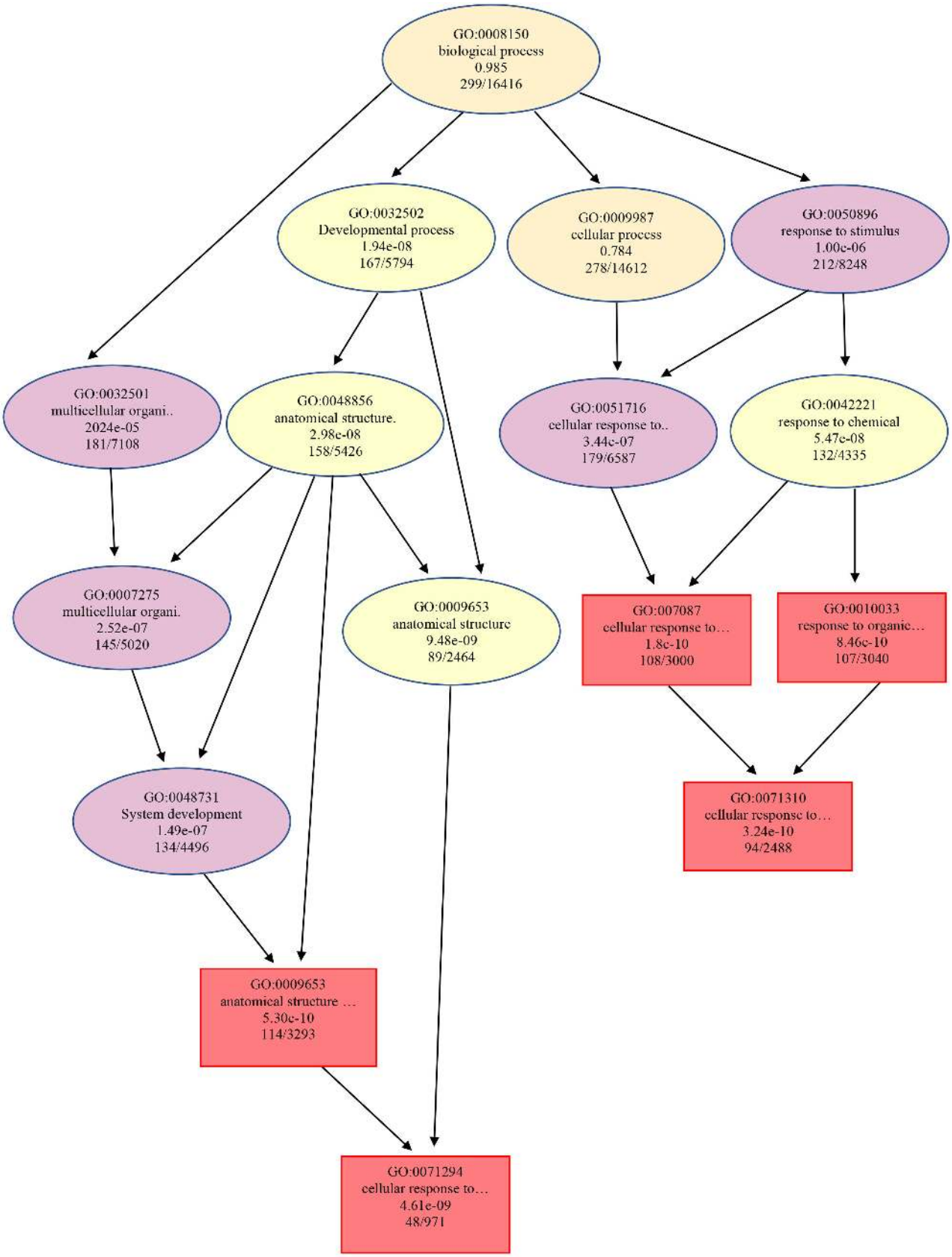


FIGURE 4. The complete DAG by means of GSEA on GSE77558 where rectangles indicate the 5 most significant terms. Red rectangle colour nodes represent the most significant terms and the remaining eliptical shaped nodes indicate least significant GO terms.

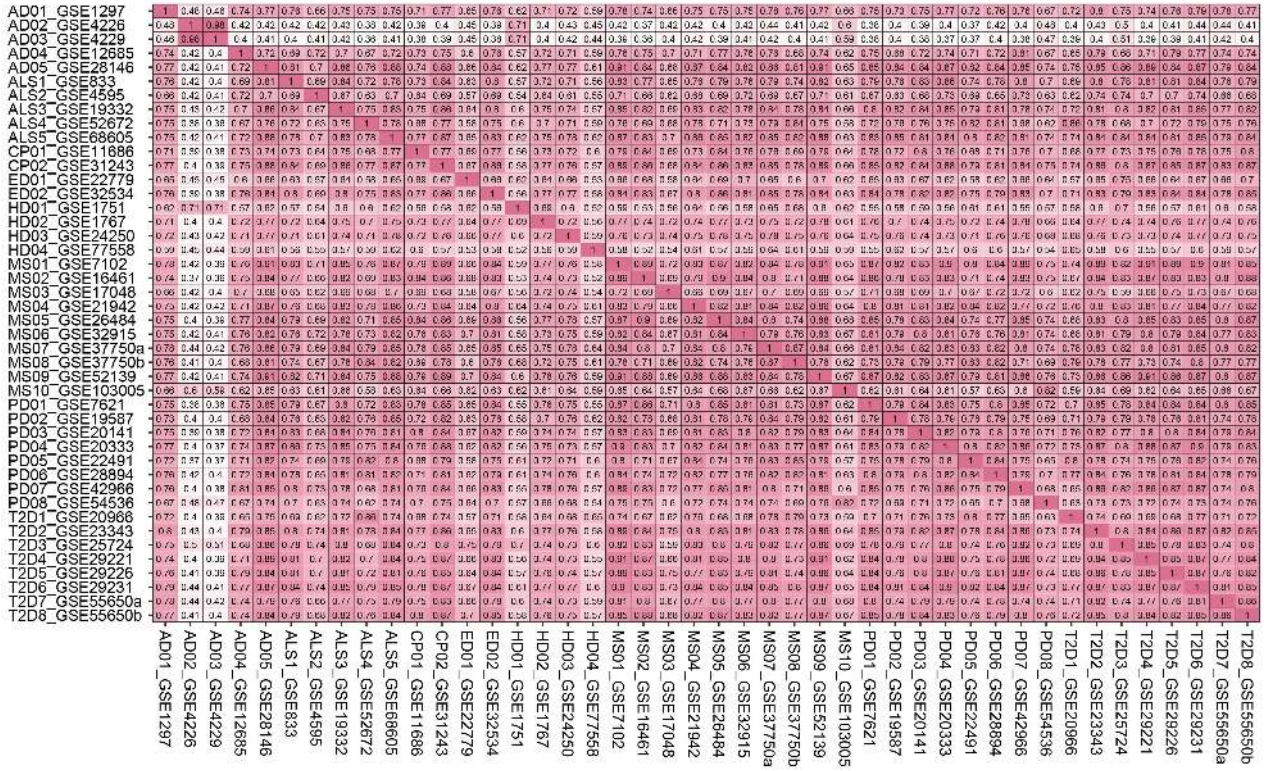


FIGURE 5. **The matrix of semantic similarity in terms of differential expressed genes from the top five GO terms.The dataset legend of the matrix is comprised of disease acronym, order, and accession number.**

exact test was applied to gene count. We identified the most significant GO terms (top 5) based on p- values by classical enrichment analysis. The DAG graph reveals that all the GO terms are non-trivial and the most significant nodes are represented as rectangles. For each node in the graph, the first line is GO ID, the second line is GO name, third is p-value and the last line is the ratio of the total number of significant genes and the total number of annotated genes to the respective GO term. Black arrows indicate is-a relationships.

However, after identifying the significant GO terms for T2D and NDs, we compared the terms. The GO terms comparison between T2D disease and NDs highlights the following in common:

- GO:0001502: cartilage condensation;
- GO:0098743: cell aggregation;
- GO:0006959: humoral immune response;
- GO:0002455: humoral immune response mediated by circulating immunoglobulin;
- GO:0006958: complement activation, classical pathway;
- GO:0007186: G protein-coupled receptor signaling pathway;
- GO:0043062: extracellular structure organization;
- GO:0030198: extracellular matrix organization;
- GO:0030199: collagen fibril organization;
- GO:0016477: cell migration;
- GO:0048870: cell motility;
- GO:0002673: regulation of acute inflammatory response;
  item GO:0030199: collagen fibril organization;
- GO:0007059: chromosome segregation;
- GO:0022610: biological adhesion;
- GO:0071294: cellular response to zinc ion;
- GO:0007632: visual behavior;
- GO:0035459: cargo loading into vesicle;
- GO:0038128: ERBB2 signaling pathway;
- GO:0009653: anatomical structure morphogenesis;
- GO:1902284: neuron projection extension involved in neuron projection guidance;
- GO:0007416: synapse assembly;
- GO:0048483: autonomic nervous system development;
- GO:0060337: type I interferon signaling pathway;
- GO:0071357: cellular response to type I interferon;
- GO:1903522: regulation of blood circulation;
- GO:0050906: detection of stimulus involved in sensory perception;
- GO:0050877: nervous system process;
- GO:0007600: sensory perception;
- GO:0010469: regulation of signaling receptor activity.

D. SEMANTIC SIMILARITY

The result of semantic similarity in terms of DEGs among pathologies is shown in Figure 5. Considering semantic similarity value above 0.6, T2D4_GSE29221 and

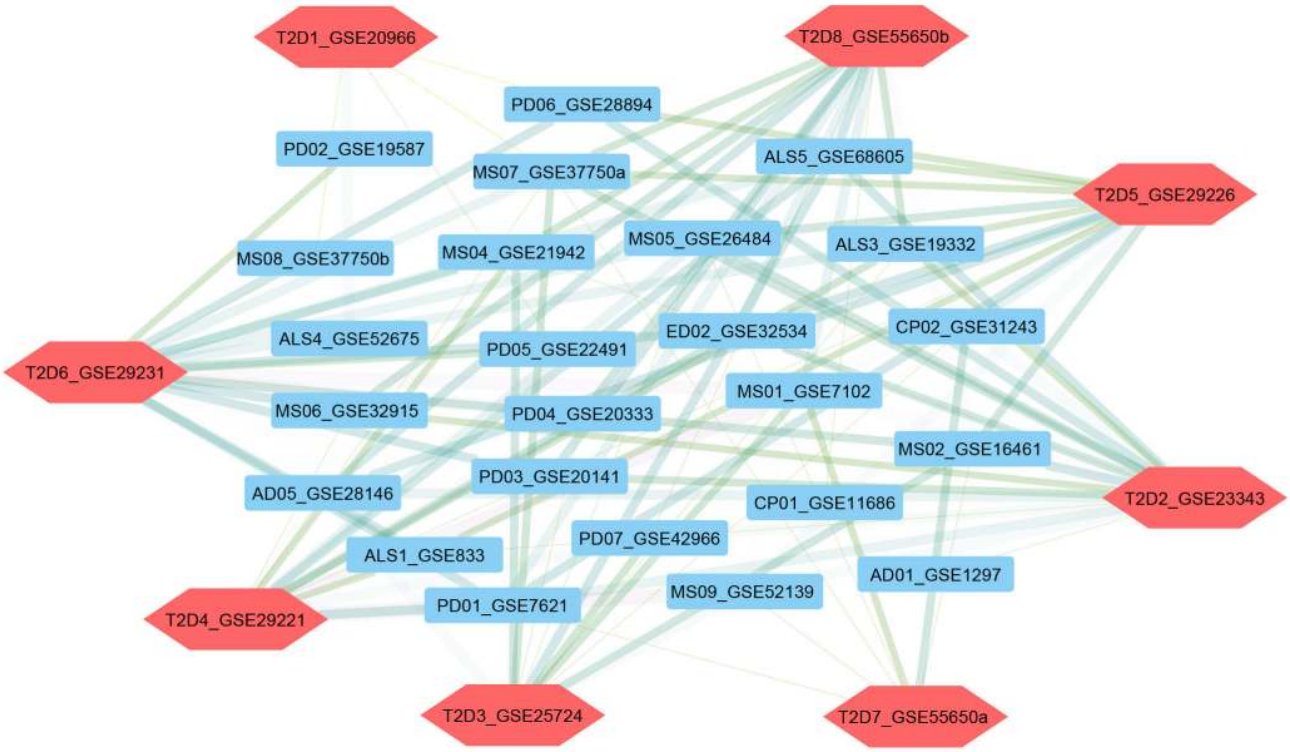


FIGURE 6. The network for representing the gene semantic similarity matrix shown in figure 5. The width of the edges is proportional to the semantic score and the red colour hexagonal node represents the T2D and the light blue colour rectangular node represents T2D associated NDs.

T2D5_GSE29226 are associated with all the selected neurological comorbidities except HD01_GSE1751 and HD04_GSE77558. Similarly, T2D1_GSE20966 is associated with all the selected neurological comorbidities except HD01_GSE1751 and ED01_GSE22779. T2D2_GSE23343 is associated with all the selected neurological comorbidities except HD04_GSE77558 whereas T2D8_GSE55650b is associated with all the selected neurological comorbidities except HD04_GSE77558. With respect to other evidence from T2D3_GSE25724, all neurological diseases are associated with T2D. In addition, AD02_GSE4226 and AD03_GSE4229 are outliers for all T2D datasets except T2D3_GSE25724 with a threshold value of absolute logFC 1.

After representing the gene semantic similarity matrix as shown in figure 5 into gene semantic similarity network as shown in Figure 6 by considering the semantic similarity threshold value of 0.77 using Cytoscape [97] we found that T2D such as T2D1_GSE20966, T2D2_GSE23343, T2D3_GSE25724, T2D4_GSE29221, T2D5_GSE29226 T2D6_GSE29231, T2D7_GSE55650a T2D8_GSE55650b are strongly associated with all the NDs such as AD, ALS, CP, ED, HD, MS, and PD.

The semantic similarity matrix of GO terms is depicted in Figure 7. Notably, T2D7_GSE55650a with AD03_GSE4229, T2D5_GSE29226 with AD05_GSE28146 and MS02_GSE16461, T2D6_GSE29231 with MS10_GSE1030 05 as well as T2D7_GSE55650a with PD02_GSE19587 have semantic similarity value of 1. If we examine semantic similarity value over 0.8, some of AD, ALS, CP, ED, HD, MS, and PD have noteworthy similarity with some of the T2D datasets. Moreover, inspecting semantic similarity value over 0.7, T2D3_GSE25724 is well clustered with all the neurological diseases except MS08_GSE37750b and ALS4_GSE52672. In addition, MS08_GSE37750b, ED02_GSE32534, ALS5_GSE68605 and AD04_GSE12685 have semantic similarity under 0.05 for T2D5_GSE29226 and T2D1_GSE29231 whereas ALS5_GSE12685 has semantic similarity under 0.05 for T2D6_GSE29231.

After representing the GO semantic similarity matrix shown in figure 7 into gene semantic similarity network as shown in Figure 8 by considering the semantic similarity threshold value of 0.80 using Cytoscape [97] we found that T2D such as T2D1_GSE20966, T2D2_GSE23343, T2D3_GSE25724, T2D4_GSE29221, T2D5_GSE29226, T2D6_GSE29231, T2D7_GSE55650a T2D8_GSE55650b are strongly associated with all the selected NDs such as AD, ALS, CP, ED, HD, MS, and PD.

We applied this proposed bioinformatics methodology to identify T2D and its neurological comorbidities on microarray gene expression datasets. Identifying the interactions among a group of diseases at the molecular level can enhance our insights into disease mechanisms. The use of semantic similarity in terms of genes and GO terms to measure the disease comorbidity enhances the identification

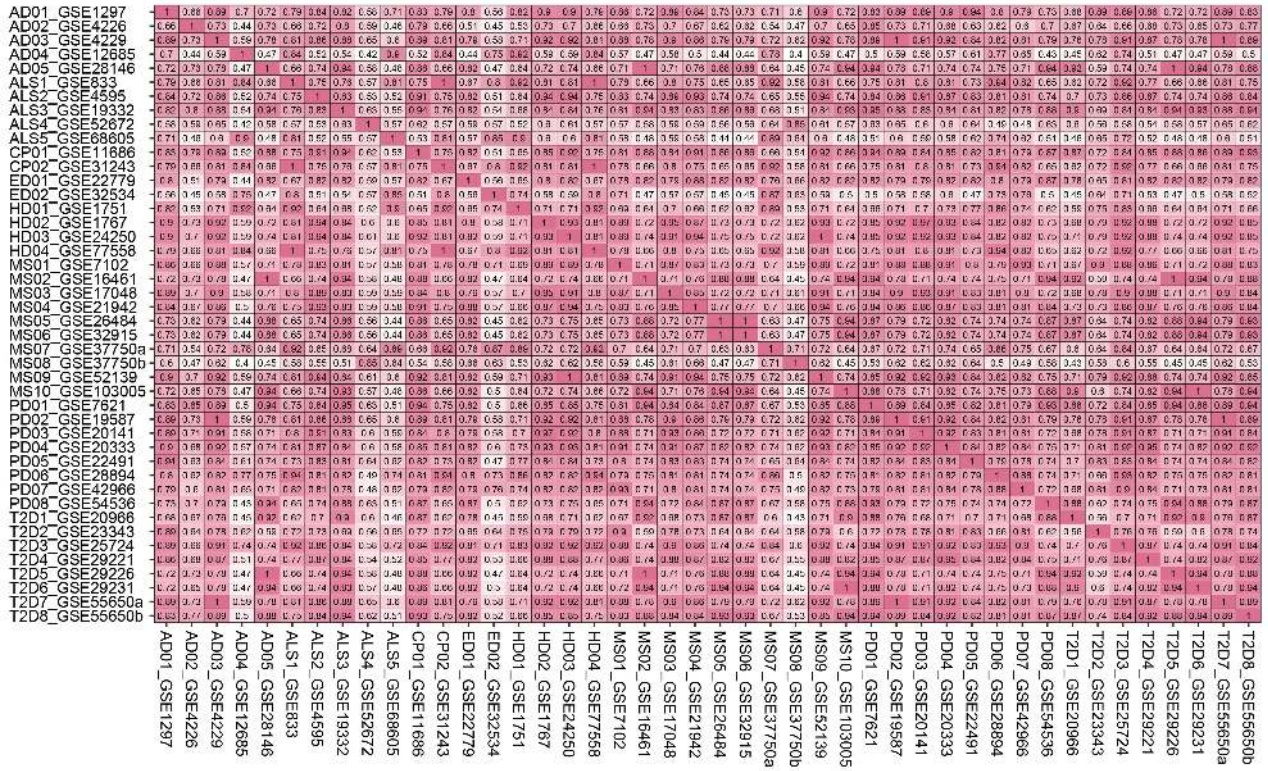


FIGURE 7. The matrix of semantic similarity is based on the top 5 GO terms. The dataset legend of the matrix is comprised of disease acronym, order, and accession number.

and characterization beyond simply identifying novel biological processes involved in each disease and eventually yields an opportunity for developing diagnosis and effective treatment. Most previous methods were designed for disease comorbidity study by considering either single omics or clinical dataset as for example comorbidity [90], [91], comoR [93], POGO [94], Comorbidity4j [95], comoRbidity [96] and CytoCom [97]. The R package 'comorbidity' is able to predict ICD-9-CM codes on the basis of comorbidity indices. This method measures the total comorbidity count or the total Charlson score [90], [91]. In [92], Hidalgo *et al.* identified comorbidity association using clinical data. In [93], the R package comoR measures relative risk and ϕ-correlation leveraging diagnosis, gene expression, and clinical data and identified related genes and pathways to predict comorbidity. This method considers only gene expression and molecular data. Moni et. al developed [94] an R software tool "POGO" to identify disease comorbidity by considering omics, phenotype, and ontological data but this method did not take into account the genetic effects on diseases. In [95], Ronzano *et al.* developed a web-based open-source software tool Comorbidity4j to identify a set of comorbidity indices using clinical data. In [96], Gutièrrez-Sacristán *et al.* have developed a method comoRbidity that performs analyses of disease comorbidity combining clinical data and genotype-phenotype based information although this method did not take into account the genetic effects on diseases. In [97], Moni *et al.* developed a tool CytoCom for Cytoscape app to visualize disease comorbidity network.

Compared to previous methods, most previously published methods in this area were designed to identify the causal relationship for disease comorbidities by considering either a single omics or clinical datasets. We have in contrast applied an integrated approach leveraging large numbers of publicly available gene expression datasets with pathway information, gene ontology data from microarray experiments which is a highly effective means of identifying pathways relevant to comorbidity interaction. The use of so many datasets from different sources and cell types maximizes the power of this approach by reducing biases of the datasets and improve the information regarding other previous studies. Our pipeline would thus be useful to reveal hidden pathological information in many different types of published gene expression datasets. As far we know, this is the first study for T2D and NDs comorbidity analysis by incorporating Gene Set Enrichment Analysis and Semantic Similarity and perhaps there are no previous comorbidity analysis methods for T2D and NDs incorporating Semantic Similarity. The methodology ensures the possibility of reusing available data and may identify disease-causing DEGs, gene function, GO terms and molecular pathways. The findings documented in this result section are likely to be improved by exploiting the maximum power of datasets from different sources and cell types. Thus, this methodology could be helpful to reveal new information from previously published datasets.

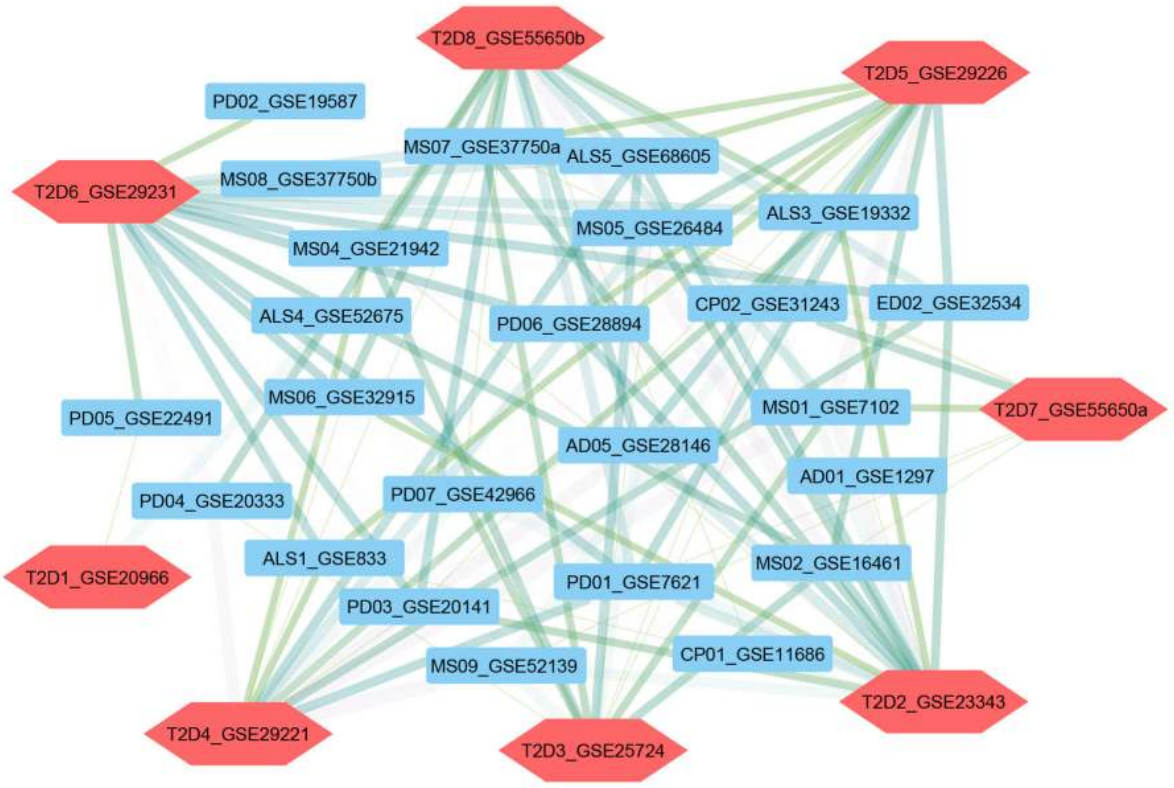


FIGURE 8. The network for representing GO semantic similarity matrix shown in figure 7. The width of the edges is proportional to the semantic score and the red colour hexagonal node represents the T2D and the light blue colour rectangular node represents T2D associated NDs.

E. POTENTIAL TARGETS VERIFICATION USING GOLD BENCHMARK DATABASES AND LITERATURE

To verify our identified potential targets, we used gold benchmark databases and an investigation of the literature. OMIM, OMIM Expanded, and dbGaP databases are called gold benchmark databases because these databases contain curated and validated genes and disease association data from the literature. In this study, we presented a combined relation of OMIM, OMIM Expanded, and dbGaP databases. For evaluating the validity of our work, we provided statistically significant differentially expressed genes of T2D to the online tool EnrichR [98] and collected statistically significant genes and their corresponding NDs from OMIM, OMIM Expanded, and dbGaP databases by choosing p-value of 0.05. The number of genes associated with each ND is educed and summarized in fewer genes with their molecular functions by checking the literature to find genes that have been clinically used as biomarkers for any of the NDs. Table 4 shows the verified potential targets and their corresponding diseases using gold benchmark databases and literature. For the verification of AD, ALS, MS, and PD, we utilized gold benchmark databases through online tool EnrichR and genes are curated in fewer genes by checking the literature. On the other hand, for the verification of CP, ED, and HD, we used literature which found genes associated with each ND. Munshi *et al.* [99] identified APP, SORL1 and CR1, Cauwenberghe *et al.* [101], [102] identified MEF2C genes associated with AD which consistent with our study. Chen *et al.* [103]–[105] identified PON2, Eykens et. al [106] identified APOE and Souza *et al.* [107], [108] identified ITPR2 and VAPB genes associated with ALS which consistent with our study. Fahey *et al.* [109] identified the TENM1 gene associated with CP and consistent with our study. Myers *et al.* [113] identified CHD2 and PURA genes associated with ED and consistent with our study. Arning and Epplen [114] reported the ITPR2 gene associated with HD. Baranzini [115] identified HLA-DQB1 and IL2RA genes, Fagerberg *et al.* [110] identified HMGA2 and Gabriele et. al identified CD24 gene associated with MS. Arning and Epplen [114] identified UCHL1, Fagerberg et. al reported TSHR, Myers *et al.* [113] identified SNCA and Lin et. al identified MAPT gene associated with

TABLE 4. Potential targets verification using gold standard benchmark databases and literatures.

Disease name	P-value	Associated Genes
AD	1.19E-03	APP; CR1; MEF2C; SORL1
ALS	2.57E-02	PON2; APOE; ITPR2; VAPB
MS	1.68E-02	HLA-DQB1; HMGA2; IL2RA; CD24
PD	7.91E-03	UCHL1; SNCA; MAPT; TSHR

PD which consistent with our study. This verification process generally confirms that significant genes that we have identified in NDs have some known disease associations.

The molecular basis of the verified potential targets are as follows: Amyloid precursor protein (APP) gene encodes a cell surface receptor and transmembrane protein that is cleaved by β- and γ- secretase in a sequential manner to yield Aβ-peptides (including Aβ40 and Aβ42). The Aβ peptides are involved in the pathogenesis of AD [99]. SORL1 (Sortilin related receptor 1) also known as SORLA and LR11 is involved in regulating protein movements through the cell and identified as biomarkers for AD [99]. Complement C3b/C4b receptor 1 (CR1) gene mediates cellular binding to particles and immune complexes that have activated complement and associated with AD [99], [100]. Myocyte enhancer factor 2C (MEF2C) plays a role in myogenesis. Mutations of this gene associated with immune response, neuronal development and synaptic function for developing AD [48], [101]. Paraoxonase 2 (PON2) is involved in the hydrolysis of lactones and the detoxification of organophosphate pesticides, neurotoxins, and aromatic esters. The mutations of PON gene polymorphisms are associated with ALS [103]. Apolipoprotein E (apoE denotes protein, APOE denotes gene) is a lipid transport protein in the brain. The mutations APOE results in a major risk of ALS [106]. Inositol 1, 4, 5-trisphosphate receptor type 2 (ITPR2) is involved in many processes including cell migration, cell division, smooth muscle contraction, and neuronal signaling. A mutation in this gene has been associated with ALS and HD [107]. Vesicle-associated membrane protein-associated protein B (VAPB) is an integral endoplasmic reticulum membrane protein, which has various functions such as intracellular vesicle trafficking, lipid transport, and the unfolded protein response and associated with ALS [107]. Teneurin transmembrane protein 1(TENM1) functions as a cellular signal transducer. Its mutation involved in neural development, regulating the establishment of proper connectivity within the nervous system causes CP [109]. C-C Motif Chemokine Ligand 2 (CCL2) is superfamily of secreted proteins involved in immunoregulatory and inflammatory processes and associated with CP [110]. Cell adhesion molecule L1-like (CHL1) plays a role in neuronal positioning of pyramidal neurons and in the regulation of both the number of interneurons and the efficacy of GABAergic synapses and associated with CP [111]. Gremlin 1 (GREM1) plays a role in regulating organogenesis, body patterning, and tissue differentiation associated with CP [110]. GATM glycine amidinotransferase (GATM) is involved in creatine biosynthesis. Mutations in this gene cause arginine: glycine amidinotransferase deficiency, an inborn error of creatine synthesis characterized by cognitive disability, language impairment, and behavioral disorders associated with ED [112]. Ubiquitin protein ligase E3A (UBE3A) is maternally expressed in the brain and bi-allelically expressed in other tissues. Maternally inherited deletion of this gene causes Angelman syndrome, characterized by severe motor and intellectual retardation, ataxia, hypotonia, epilepsy, absence of speech, and characteristic facies [112]. CHD2 (Chromodomain Helicase DNA Binding Protein 2) is a protein-coding gene. Its function may be regulation of chromatin structure, and diseases associated with CHD2 include epileptic encephalopathy [113]. Purine rich element binding protein A(PURA) is involved in the formation or maturation of myelin, the protective substance that covers nerves and promotes the efficient transmission of nerve impulses and associated with ED [113]. Ubiquitin carboxy-terminal hydrolase L1 (UCHL1) is a de-ubiquitinating enzyme with important functions in the recycling of ubiquitin expressed in the neurons and cells of the diffuse neuroendocrine system and associated with HD [114]. The protein encoded by MATR3 (Matrin 3) gene is localized in the nuclear matrix plays a role in transcription or may interact with other nuclear matrix proteins to form the internal fibro granular network and is associated with HD [110]. C9orf72-SMCR8 complex subunit (C9orf72) gene provides instructions for making a protein that is abundant in nerve cells (neurons) in the outer layers of the brain (cerebral cortex) and in specialized neurons in the brain and spinal cord that control movement (motor neurons) and associated with HD [110]. HLA-DQB1 major histocompatibility complex, class II, DQ beta 1(HLA-DQB1) belongs to the HLA class II beta chain paralogs and plays a central role in the immune system by presenting peptides derived from extracellular proteins and associated with MS [115]. High mobility group AT-hook 2 (HMGA2) functions as a transcriptional regulator and associated with MS [110]. Interleukin 2 receptor subunit alpha (IL2RA) is involved in the regulation of immune tolerance by controlling regulatory T cells (TREGs) activity and associated with MS [115]. The protein encoded by signal transducer CD24 also known as cluster of differentiation 24 or heat-stable antigen CD24 (HSA) contributes to a wide range of downstream signaling networks and is crucial for neural development and associated with MS [116], [117]. Ubiquitin carboxy-terminal hydrolase L1 (UCHL1) gene is a de-ubiquitinating enzyme with important functions in the recycling of ubiquitin and is specifically expressed in the neurons and in cells of the diffuse neuroendocrine system associated with PD [114]. Alpha-synuclein (SNCA) protein encoded by SNCA (Alpha-synuclein) is present at high levels in neurons and regulates dopamine neurotransmission, and its aberrant expression is associated with PD [118]. The protein encoded by the thyroid-stimulating hormone receptor (TSHR) gene is a membrane protein and a major controller of thyroid cell metabolism. Defects in this gene are also seen in PD [110]. Microtubule-associated protein tau (MAPT) encodes a microtubule-associated protein tau, which has a role in stabilizing microtubules in the neurons and its mutations are associated with PD [119].

IV. DISCUSSION

The goal of this research is to establish the ability of an integrated pipeline of bioinformatics methodology to extract discriminative information from public data repositories and identifying the relationships of complex

diseases such as T2D and its neurological comorbidities. We applied this proposed bioinformatics methodology to identify T2D and its neurological comorbidities on microarray data from Gene Expression Omnibus (GEO) repository: (https://www.ncbi.nlm.nih.gov/geo).

We employed GSEA to study T2D with regard to DEGs, molecular pathways, and interrelationship among omics data such as Gene Ontology. A particular feature of our proposed methodology is computing the proximity (in terms of semantic similarity) among different datasets based on their selected ontology. Although we applied our proposed integrated bioinformatics pipeline for T2D and its neurological disease comorbidities, it is methodically universal and can be applied for other diseases and their comorbidities along with more complex pathologies. Our proposed pipeline shows a novel use of freely available research data in exploring disease comorbidity from a bioinformatics point of view. The only manual task needed to be was that rather than the automatic selection of GEO samples, we reviewed the GSM records manually and classified samples into sick, healthy or treated and created design models.

At least 3 disease-affected and 3 healthy control samples were considered in our study for each disease case to exploit the maximum power of the study. We began from the set of DEGs and carried out GSEA on them to identify the top 5 GO terms. A comparison among all the selected pathologies in terms of genes and GO terms were computed by applying the method of semantic similarity. The use of semantic similarity in terms of genes and GO terms to measure the disease comorbidity enhances the identification and characterization beyond simply identifying novel biological processes involved in each disease. Our proposed integrated bioinformatics methodology is implemented in programming language R by incorporating several Bioconductor packages. Using the threshold p-value of 0.05 and absolute logFC value of 1, we identified discriminative sets of DEGs and GO terms which improve on current approaches to finding common DEGs and biological pathways among a group of diseases of interest. We have applied the process on microarray datasets for the selected pathologies that are publicly available and also suit our approach.

The proposed methodology is data-driven and it should be noted that the choice of the dataset would have qualitative and quantitative impacts on the results, and the use of a larger number of microarray datasets may enrich the evidence. We observe that taking into account data from different sources and cell types could strengthen the evidence obtained. Our proposed methodology used here for T2D and neurological comorbidities could be widely used, and it has two particular uses to uncover possible mechanisms of T2D-associated activities that drive the development of neurological diseases, and as a means to identify possible significant comorbidities.

The development of bioinformatics methodology by utilizing omics and molecular data is providing new opportunities for medical practitioners to enhance clinical decision-making such as disease risk evaluation, disease diagnosis, and subtyping, drug therapy, and dose selection [120] and represent a step toward the development of truly personalized medicine. Thus, our methodology can provide fundamental new insights into disease mechanisms, and such identified disease mechanisms could be useful for further investigations to develop novel therapeutic targets.

V. CONCLUSION

In this study, we have considered transcriptomics, omics and molecular level data to investigate how the proposed methodology can be used to identify neurological disease comorbidities with T2D. We found that these neurological diseases are highly connected with T2D in terms of common biological processes, pathways and omics data as for example GO. Our findings indicate that the progression of complex diseases could be identified and studied using bioinformatics methodology as it offers the potential to enhance our understanding of complex human diseases. Identification of comorbidity interactions has a clinical interest because it may reveal new information about disease-causing factors as well as new therapeutic targets. This study demonstrates the worth of an integrated bioinformatics methodology in revealing possible disease relationships and opportunities for drug repositioning. Thus, we can say that this kind of approach will be helpful for making evidence-based recommendations about disease comorbidities. Our proposed approach could be extended as a comorbidity map by integrating other disease data besides neurological diseases. Our methodology is also likely to be useful not only for T2D research but also for the study of other complex diseases. Researchers and medical practitioners may use it as an important tool to uncover details of underlying disease mechanisms that underpin the biology and etiology of disease comorbidity and for the establishment of more effective and efficient treatments, perhaps within the context of an individualized and personalized pharmacotherapy.

REFERENCES

[1] D. Premilovac, R. J. Gasperini, S. Sawyer, A. West, M. A. Keske, B. V. Taylor, and L. Foa, "A new method for targeted and sustained induction of type 2 diabetes in rodents," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 14158.

[2] M. Y. Donath and S. E. Shoelson, "Type 2 diabetes as an inflammatory disease," *Nature Rev. Immunol.*, vol. 11, no. 2, p. 98, 2011.

[3] M. N. Haidar, M. B. Islam, U. N. Chowdhury, M. R. Rahman, F. Huq, J. M. Quinn, and M. A. Moni, "Network-based computational approach to identify genetic links between cardiomyopathy and its risk factors," *IET Syst. Biol.*, Oct. 2019, doi: 10.1049/iet-syb.2019.0074.

[4] S. Bonner-Weir, D. F. Trent, and G. C. Weir, "Partial pancreatectomy in the rat and subsequent defect in glucose-induced insulin release," *J. Clin. Invest.*, vol. 71, no. 6, pp. 1544–1553, 1983, doi: 10.1172/jci110910.

[5] J. L. Leahy, H. E. Cooper, D. A. Deal, and G. C. Weir, "Chronic hyperglycemia is associated with impaired glucose influence on insulin secretion. A study in normal rats using chronic *in vivo* glucose infusions," *J. Clin. Invest.*, vol. 77, no. 3, pp. 908–915, Mar. 1986.

[6] M. Y. Donath, D. J. Gross, E. Cerasi, and N. Kaiser, "Hyperglycemia-induced β-cell apoptosis in pancreatic islets of psammomys obesus during development of diabetes," *Diabetes*, vol. 48, no. 4, pp. 738–744, Apr. 1999.

[7] G. C. Weir and S. Bonner-Weir, "Five stages of evolving β-cell dysfunction during progression to diabetes," *Diabetes*, vol. 53, no. suppl 3, pp. S16–S21, Dec. 2004.

[8] L. Rossetti, D. Smith, G. I. Shulman, D. Papachristou, and R. A. DeFronzo, "Correction of hyperglycemia with phlorizin normalizes tissue sensitivity to insulin in diabetic rats," *J. Clin. Invest.*, vol. 79, no. 5, pp. 1510–1515, May 1987.

[9] G. M. Reaven, C. Hollenbeck, C. Y. Jeng, M. S. Wu, and Y. D. Chen, "Measurement of plasma glucose, free fatty acid, lactate, and insulin for 24 h in patients with NIDDM," *Diabetes*, vol. 37, no. 8, pp. 1020–1024, 1988.

[10] K. Z. Walker, K. O'Dea, L. Johnson, A. J. Sinclair, L. S. Piers, G. C. Nicholson, and J. G. Muir, "Body fat distribution and non-insulin dependent diabetes: Comparison of a fiber-rich, high carbohydrate, low-fat (23%) diet and a 35% fat diet high in monounsaturated fat," *Amer. J. Clin. Nutrition*, vol. 63, no. 2, pp. 254–260, Feb. 1996.

[11] K. Maedler, J. Oberholzer, P. Bucher, G. A. Spinas, and M. Y. Donath, "Monounsaturated fatty acids prevent the deleterious effects of palmitate and high glucose on human pancreatic β-cell turnover and function," *Diabetes*, vol. 52, no. 3, pp. 726–733, Mar. 2003.

[12] K. Maedler, G. A. Spinas, D. Dyntar, W. Moritz, N. Kaiser, and M. Y. Donath, "Distinct effects of saturated and monounsaturated fatty acids on β-cell turnover and function," *Diabetes*, vol. 50, no. 1, pp. 69–76, Jan. 2001.

[13] V. Poitout and R. P. Robertson, "Glucolipotoxicity: Fuel excess and β-cell dysfunction," *Endocrine Rev.*, vol. 29, no. 3, pp. 351–366, Nov. 2007.

[14] J. L. Evans, I. D. Goldfine, B. A. Maddux, and G. M. Grodsky, "Oxidative stress and stress-activated signaling pathways: A unifying hypothesis of type 2 diabetes," *Endocrine Rev.*, vol. 23, no. 5, pp. 599–622, Oct. 2002.

[15] J. L. Evans, I. D. Goldfine, B. A. Maddux, and G. M. Grodsky, "Are oxidative stress-activated signaling pathways mediators of insulin resistance and β-cell dysfunction?" *Diabetes*, vol. 52, pp. 1–8, Jan. 2003.

[16] H. P. Harding and D. Ron, "Endoplasmic reticulum stress and the development of diabetes: A review," *Diabetes*, vol. 51, no. suppl 3, pp. S455–S461, 2002.

[17] E. Araki, S. Oyadomari, and M. Mori, "Endoplasmic reticulum stress and diabetes mellitus," *Internal Med.*, vol. 42, no. 1, pp. 7–14, Jan. 2003.

[18] T. Izumi, H. Yokota-Hashimoto, S. Zhao, J. Wang, P. A. Halban, and T. Takeuchi, "Dominant negative pathogenesis by mutant proinsulin in the Akita diabetic mouse," *Diabetes*, vol. 52, no. 2, pp. 409–416, Feb. 2003.

[19] G. S. Hotamisligil, "Endoplasmic reticulum stress and the inflammatory basis of metabolic disease," *Cell*, vol. 140, no. 6, pp. 900–917, Mar. 2010.

[20] S. Zraika, R. L. Hull, C. B. Verchere, A. Clark, K. J. Potter, P. E. Fraser, D. P. Raleigh, and S. E. Kahn, "Toxic oligomers and islet β cell death: Guilty by association or convicted by circumstantial evidence?" *Diabetologia*, vol. 53, no. 6, pp. 1046–1056, Jun. 2010.

[21] O. Ali, "Genetics of type 2 diabetes," *World J. Diabetes*, vol. 4, no. 4, p. 114, Aug. 2013.

[22] B. F. Voight, L. J. Scott, V. Steinthorsdottir, A. P. Morris, C. Dina, and R. P. Welch, E. Zeggini, C. Huth, Y. S. Aulchenko, and G. Thorleifsson, "Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis," *Nature Genet.*, vol. 42, no. 7, p. 579, 2010.

[23] R. A. Hackett and A. Steptoe, "Type 2 diabetes mellitus and psychological stress—A modifiable risk factor," *Nature Rev. Endocrinol.*, vol. 13, no. 9, p. 547, 2017.

[24] A. Karamitri and R. Jockers, "Melatonin in type 2 diabetes mellitus and obesity," *Nature Rev. Endocrinol.*, vol. 15, no. 2, pp. 105–125, Feb. 2019.

[25] E. Lalla and P. N. Papapanou, "Diabetes mellitus and periodontitis: A tale of two common interrelated diseases," *Nature Rev. Endocrinol.*, vol. 7, no. 12, p. 738, 2011.

[26] A. P. Kong, G. Xu, N. Brown, W. Y. So, R. C. Ma, and J. C. Chan, "Diabetes and its comorbidities—Where East meets West," *Nature Rev. Endocrinology*, vol. 9, no. 9, p. 537, 2013.

[27] R. Pradeepa and V. Mohan, "Prevalence of type 2 diabetes and its complications in India and economic costs to the nation," *Eur. J. Clin. Nutrition*, vol. 71, no. 7, p. 816, 2017.

[28] K. C. Paul, M. Jerrett, and B. Ritz, "Type 2 diabetes mellitus and Alzheimer's disease: Overlapping biologic mechanisms and environmental risk factors," *Current Environ. Health Rep.*, vol. 5, no. 1, pp. 44–58, 2018.

[29] G. Hu, P. Jousilahti, S. Bidel, R. Antikainen, and J. Tuomilehto, "Type 2 diabetes and the risk of Parkinsons disease," *Diabetes Care*, vol. 30, no. 4, pp. 842–847, Apr. 2007.

[30] D. Mariosa, F. Kamel, R. Bellocco, W. Ye, and F. Fang, "Association between diabetes and amyotrophic lateral sclerosis in Sweden," *Eur. J. Neurol.*, vol. 22, no. 11, 2015, Art. no. 14361442.

[31] J. J. Palop and L. Mucke, "Epilepsy and cognitive impairments in Alzheimer disease," *Arch. Neurol.*, vol. 66, no. 4, 2009, Art. no. 435440.

[32] *Daily Diabetes, Type 2 Diabetes May Raise Risk of Epilepsy*. Accessed: Jun. 20, 2019. [Online]. Available: https://www.diabetesdaily.com/blog/study-type-2-diabetes-may-raise-risk-of-epilepsy-563076/

[33] N. M. Lalić, J. Marić, M. Svetel, A. Jotić, E. Stefanova, K. Lalić, N. Dragašević, T. Mili ić, L. Lukić, and V. S. Kostić, "Glucose homeostasis in Huntington disease: Abnormalities in insulin sensitivity and early-phase insulin secretion," *Arch. Neurol.*, vol. 65, no. 4, 2008, Art. no. 476480.

[34] *Guidance Cerebral Palsy, Cerebral Palsy and Diabetes*. Accessed: Jun. 20, 2019. [Online]. Available: https://www.cerebralpalsyguidance.com/cerebral-palsy/associateddisorders/diabetes/

[35] M. D. C. Janik, T. B. Newman, Y. W. Cheng, G. Xing, W. M. Gilbert, and Y. W. Wu, "Maternal diagnosis of obesity and risk of cerebral palsy in the child," *J. Paediatrics*, vol. 163, no. 5, 2013, Art. no. 13071312.

[36] W. H. Hou, C. Y. Li, H. H. Chang, Y. Sun, and C. C. Tsai, "A population-based cohort study suggests an increased risk of multiple sclerosis incidence in patients with type 2 diabetes mellitus," *J. Epidemiol.*, vol. 27, no. 5, 2017, Art. no. 235241.

[37] E. Capobianco and P. Liò, "Comorbidity: A multidimensional approach," *Trends Mol. Med.*, vol. 19, no. 9, pp. 515–521, 2013.

[38] S. E. Wilhite and T. Barrett, "Strategies to explore functional genomics data sets in NCBI's GEO database," in *Next Generation Microarray Bioinformatics*. Totowa, NJ, USA: Humana Press, Jan. 2012, pp. 41–53, doi: 10.1007/978-1-61779-400-1_3.

[39] L. Marselli, J. Thorne, S. Dahiya, D. C. Sgroi, A. Sharma, S. Bonner-Weir, P. Marchetti, and G. C. Weir, "Gene expression profiles of Beta-cell enriched tissue obtained by laser capture microdissection from subjects with type 2 diabetes," *PLoS ONE*, vol. 5, no. 7, Jul. 2010, Art. no. e11499.

[40] H. Misu *et al.*, "A liver-derived secretory protein, selenoprotein P, causes insulin resistance," *Cell Metabolism*, vol. 12, no. 5, pp. 483–495, Nov. 2010.

[41] V. Dominguez, C. Raimondi, S. Somanath, M. Bugliani, M. K. Loder, C. E. Edling, N. Divecha, G. da Silva-Xavier, L. Marselli, S. J. Persaud, M. D. Turner, G. A. Rutter, P. Marchetti, M. Falasca, and T. Maffucci, "Class II phosphoinositide 3-kinase regulates exocytosis of insulin granules in pancreatic beta cells," *J. Biol. Chem.*, vol. 286, no. 6, pp. 4216–4225, Feb. 2011.

[42] P. Jain, S. Vig, M. Datta, D. Jindel, A. K. Mathur, S. K. Mathur, and A. Sharma, "Systems biology approach reveals genome to phenome correlation in type 2 diabetes," *PLoS ONE*, vol. 8, no. 1, 2013, Art. no. e53522.

[43] N. Sakib, U. N. Chowdhury, M. B. Islam, J. M. Quinn, and M. A. Moni, "A system biology approach to identify the genetic markers to the progression of parkinson's disease for aging, lifestyle and type 2 diabetes," *bioRxiv*, Jan. 2018, Art. no. 482760.

[44] U. N. Chowdhury, M. B. Islam, S. Ahmad, F. Huq, J. M. Quinn, and M. A. Moni, "Network-based identification of genetic factors in ageing, lifestyle and type 2 diabetes that influence in the progression of Alzheimer's disease," *bioRxiv*, Jan. 2018, Art. no. 482844.

[45] A. E. Brown, J. Palsgaard, R. Borup, P. Avery, D. A. Gunn, P. De Meyts, S. J. Yeaman, and M. Walker, "P38 MAPK activation upregulates proinflammatory pathways in skeletal muscle cells from insulin-resistant type 2 diabetic patients," *Amer. J. Physiol.-Endocrinol. Metabolism*, vol. 308, no. 1, pp. E63–E70, Jan. 2015.

[46] E. M. Blalock, J. W. Geddes, K. C. Chen, N. M. Porter, W. R. Markesbery, and P. W. Landfield, "Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 7, pp. 2173–2178, Feb. 2004.

[47] O. C. Maes, H. M. Schipper, H. M. Chertkow, and E. Wang, "Methodology for discovery of Alzheimer's disease blood-based biomarkers," *J. Gerontol. A, Biomed. Sci. Med. Sci.*, vol. 64, no. 6, pp. 636–645, Jun. 2009.

[48] M. R. Rahman, T. Islam, T. Zaman, M. Shahjaman, M. R. Karim, F. Huq, J. M. Quinn, R. D. Holsinger, E. Gov, and M. A. Moni, "Identification of molecular signatures and pathways to identify novel therapeutic targets in Alzheimer's disease: Insights from a systems biomedicine perspective," *Genomics*, to be published.

[49] C. Williams, R. M. Shai, Y. Wu, Y.-H. Hsu, T. Sitzer, B. Spann, C. McCleary, Y. Mo, and C. A. Miller, "Transcriptome analysis of synaptoneurosomes identifies neuroplasticity genes overexpressed in incipient Alzheimer's disease," *PLoS ONE*, vol. 4, no. 3, 2009, Art. no. e4936.

[50] E. M. Blalock, H. M. Buechel, J. Popovic, J. W. Geddes, and P. W. Landfield, "Microarray analyses of laser-captured hippocampus reveal distinct gray and white matter signatures associated with incipient Alzheimer's disease," *J. Chem. Neuroanatomy*, vol. 42, no. 2, pp. 118–126, Oct. 2011.

[51] T. G. Lesnick, S. Papapetropoulos, D. C. Mash, J. Ffrench-Mullen, L. Shehadeh, M. De Andrade, J. R. Henley, W. A. Rocca, J. E. Ahlskog, and D. M. Maraganore, "A genomic pathway approach to a complex disease: Axon guidance and Parkinson disease," *PLoS Genet.*, vol. 3, no. 6, Jun. 2007, Art. no. e98.

[52] N. M. Lewandowski, S. Ju, M. Verbitsky, B. Ross, M. L. Geddie, E. Rockenstein, A. Adame, A. Muhammad, J. P. Vonsattel, D. Ringe, L. Cote, S. Lindquist, E. Masliah, G. A. Petsko, K. Marder, L. N. Clark, and S. A. Small, "Polyamine pathway contributes to the pathogenesis of Parkinson disease," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 39, pp. 16970–16975, Sep. 2010.

[53] B. Zheng *et al.*, "PGC-1α, a potential therapeutic target for early intervention in Parkinson's disease," *Sci. Transl. Med.*, vol. 2, no. 52, Oct. 2010, Art. no. 52ra73.

[54] E. Mutez, L. Larvor, F. Leprêtre, V. Mouroux, D. Hamalek, J.-P. Kerckaert, J. Pérez-Tur, N. Waucquier, C. Vanbesien-Mailliot, A. Duflot, D. Devos, L. Defebvre, A. Kreisler, B. Frigard, A. Destée, and M.-C. Chartier-Harlin, "Transcriptional profile of Parkinson blood mononuclear cells with LRRK2 mutation," *Neurobiol. Aging*, vol. 32, no. 10, pp. 1839–1848, Oct. 2011.

[55] A. K. Alieva, M. I. Shadrina, E. V. Filatova, A. V. Karabanov, S. N. Illarioshkin, S. A. Limborska, and P. A. Slominsky, "Involvement of endocytosis and alternative splicing in the formation of the pathological process in the early stages of Parkinson's disease," *Biomed. Res. Int.*, vol. 2014, Apr. 2014, Art. no. 718732, doi: 10.1155/2014/718732.

[56] M. A. Moni, H. K. Rana, M. B. Islam, M. B. Ahmed, H. Xu, M. A. Hasan, Y. Lei, and J. M. Quinn, "A computational approach to identify blood cell-expressed Parkinson's disease biomarkers that are coordinately expressed in brain tissue," *Comput. Biol. Med.*, vol. 113, Oct. 2019, Art. no. 103385.

[57] F. Dangond, D. Hwang, S. Camelo, P. Pasinelli, M. P. Frosch, G. Stephanopoulos, G. Stephanopoulos, R. H. Brown, Jr., and S. R. Gullans, "Molecular signature of late-stage human ALS revealed by expression profiling of postmortem spinal cord gray matter," *Physiol. Genomics*, vol. 16, no. 2, pp. 229–239, Jan. 2004.

[58] C. W. Lederer, A. Torrisi, M. Pantelidou, N. Santama, and S. Cavallaro, "Pathways and genes differentially expressed in the motor cortex of patients with sporadic amyotrophic lateral sclerosis," *BMC Genomics*, vol. 8, p. 26, Jan. 2007.

[59] L. E. Cox, L. Ferraiuolo, E. F. Goodall, P. R. Heath, A. Higginbottom, H. Mortiboys, H. C. Hollinger, J. A. Hartley, A. Brockington, C. E. Burness, and K. E. Morrison, "Mutations in CHMP2B in lower motor neuron predominant amyotrophic lateral sclerosis (ALS)," *PLoS ONE*, vol. 5, no. 3, Mar. 2010, Art. no. e9872.

[60] O. Butovsky, M. P. Jedrychowski, R. Cialic, S. Krasemann, G. Murugaiyan, Z. Fanek, D. J. Greco, P. M. Wu, C. E. Doykan, O. Kiner, R. J. Lawson, M. P. Frosch, N. Pochet, R. El Fatimy, A. M. Krichevsky, S. P. Gygi, H. Lassmann, J. Berry, M. E. Cudkowicz, and H. L. Weiner, "Targeting miR-155 restores abnormal microglia and attenuates disease in SOD1 mice," *Ann. Neurol.*, vol. 77, no. 1, pp. 75–99, Jan. 2015.

[61] J. Cooper-Knock, J. J. Bury, P. R. Heath, M. Wyles, A. Higginbottom, C. Gelsthorpe, J. R. Highley, G. Hautbergue, M. Rattray, J. Kirby, and P. J. Shaw, "C9ORF72 GGGGCC expanded repeats produce splicing dysregulation which correlates with disease severity in amyotrophic lateral sclerosis," *PLoS ONE*, vol. 10, no. 5, 2015, Art. no. e0127376.

[62] M. Carlet, K. Janjetovic, J. Rainer, S. Schmidt, R. Panzer-Grümayer, G. Mann, M. Prelog, B. Meister, C. Ploner, and R. Kofler, "Expression, regulation and function of phosphofructo-kinase/fructose-biphosphatases (PFKFBs) in glucocorticoid-induced apoptosis of acute lymphoblastic leukemia cells," *BMC Cancer*, vol. 10, p. 638, Nov. 2010.

[63] C. E. Niesen, J. Xu, X. Fan, X. Li, C. J. Wheeler, A. N. Mamelak, and C. Wang, "Transcriptomic profiling of human peritumoral neocortex tissues revealed genes possibly involved in tumor-induced epilepsy," *PLoS ONE*, vol. 8, no. 2, 2013, Art. no. e56077.

[64] F. Borovecki, L. Lovrecic, J. Zhou, H. Jeong, F. Then, H. D. Rosas, S. M. Hersch, P. Hogarth, B. Bouzou, R. V. Jensen, and D. Krainc, "Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 31, pp. 11023–11028, Aug. 2005.

[65] M. H. Rahman, S. Peng, C. Chen, P. Lio, and M. A. Moni, "Genetic effect of type 2 diabetes to the progression of neurological diseases," *bioRxiv*, Jan. 2018, Art. no. 480400.

[66] Y. Hu, V. Chopra, R. Chopra, J. J. Locascio, Z. Liao, H. Ding, B. Zheng, W. R. Matson, R. J. Ferrante, H. D. Rosas, S. M. Hersch, and C. R. Scherzer, "Transcriptional modulator H2A histone family, member Y (H2AFY) marks Huntington disease activity in man and mouse," *Proc. Nat. Acad. Sci. USA*, vol. 108, no. 41, pp. 17141–17146, Oct. 2011.

[67] E. D. Nekrasov, V. A. Vigont, S. A. Klyushnikov, O. S. Lebedeva, E. M. Vassina, A. N. Bogomazova, I. V. Chestkov, T. A. Semashko, E. Kiseleva, L. A. Suldina, P. A. Bobrovsky, O. A. Zimina, M. A. Ryazantseva, A. Y. Skopin, S. N. Illarioshkin, E. V. Kaznacheyeva, M. A. Lagarkova, and S. L. Kiselev, "Manifestation of Huntington's disease pathology in human induced pluripotent stem cell-derived neurons," *Mol. Neurodegener.* vol. 11, p. 27, Apr. 2016.

[68] L. R. Smith, E. Pontén, Y. Hedström, S. R. Ward, H. G. Chambers, S. Subramaniam, and R. L. Lieber, "Novel transcriptional profile in wrist muscles from cerebral palsy patients," *BMC Med. Genomics*, vol. 2, p. 44, Jul. 2009.

[69] L. R. Smith, H. G. Chambers, S. Subramaniam, and R. L. Lieber, "Transcriptional abnormalities of hamstring muscle contractures in children with cerebral palsy," *PLoS ONE*, vol. 7, no. 8, 2012, Art. no. e40686.

[70] V. Annibali, G. Ristori, D. F. Angelini, and B. Serafini, "CD161^{high} CD8^{+} T cells bear pathogenetic potential in multiple sclerosis," *Brain*, vol. 134, pp. 542–554, Feb. 2011.

[71] C. Riveros, D. Mellor, K. S. Gandhi, F. C. McKay, M. B. Cox, R. Berretta, S. Y. Vaezpour, M. Inostroza-Ponta, S. A. Broadley, R. N. Heard, S. Vucic, G. J. Stewart, D. W. Williams, R. J. Scott, J. Lechner-Scott, D. R. Booth, and P. Moscato, "A transcription factor map as revealed by a genome-wide gene expression analysis of whole-blood mRNA transcriptome in multiple sclerosis," *PLoS ONE*, vol. 5, no. 12, Dec. 2010, Art. no. e14176.

[72] A. K. Kemppinen, J. Kaprio, A. Palotie, and J. Saarela, "Systematic review of genome-wide expression studies in multiple sclerosis," *BMJ Open*, vol. 1, no. 1, Jul. 2011, Art. no. e000053.

[73] Y. Nakatsuji *et al.*, "Elevation of Sema4A implicates Th cell skewing and the efficacy of IFN-β therapy in multiple sclerosis," *J. Immunol.*, vol. 188, no. 10, pp. 4858–4865, May 2012.

[74] T. Zrzavy, S. Hametner, I. Wimmer, O. Butovsky, H. L. Weiner, and H. Lassmann, "Loss of 'homeostatic' microglia and patterns of their activation in active multiple sclerosis," *Brain*, vol. 140, no. 7, pp. 1900–1913, Jul. 2017.

[75] L. L. Aung, A. Brooks, S. A. Greenberg, M. L. Rosenberg, S. Dhib-Jalbut, and K. E. Balashov, "Multiple sclerosis-linked and interferon-beta-regulated gene expression in plasmacytoid dendritic cells," *J. Neuroimmunol.*, vol. 250, nos. 1–2, pp. 99–105, Sep. 2012.

[76] A. Lieury, M. Chanal, G. Androdias, R. Reynolds, S. Cavagna, P. Giraudon, C. Confavreux, and S. Nataf, "Tissue remodeling in periplaque regions of multiple sclerosis spinal cord lesions," *Glia*, vol. 62, no. 10, pp. 1645–1658, Oct. 2014.

[77] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 43 pp. 15545–15550, 2005.

[78] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "KEGG as a reference resource for gene and protein annotation," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D457–D462, Oct. 2015.

[79] Gene Ontology Consortium, "Gene ontology consortium: Going forward," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D1049–D1056, Nov. 2014.

[80] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, "A new method to measure the semantic similarity of GO terms," *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, 2007.

[81] C. Pesquita, D. Faria, H. Bastos, A. E. Ferreira, A. O. Falcão, and F. M. Couto, "Metrics for GO based protein semantic similarity: A systematic evaluation," *BMC Bioinf.*, vol. 9, no. 5, p. S4, 2008.

[82] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "Limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Res.*, vol. 43, no. 7, p. e47, 2015.
[83] J. R. Alexa, "TopGO: Enrichment analysis for gene ontology, version 2.30.0," Tech. Rep., 2016, doi: 10.18129/B9.bioc.topGO.
[84] W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, and R. Gottardo, "Orchestrating high-throughput genomic analysis with Bioconductor," *Nature Methods*, vol. 12, no. 2, pp. 115–121, 2015.
[85] S. Davis and P. S. Meltzer, "GEOquery: A bridge between the gene expression omnibus (GEO) and bioconductor," *Bioinformatics*, vol. 14, pp. 1846–1847, Jul. 2007.
[86] V. C. Gentleman, W. Huber, and F. Hahne, "Genefilter: Methods for filtering genes from high-throughput experiments, version 1.60.0," Tech. Rep., 2017, doi: 10.18129/B9.bioc.genefilter.
[87] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang, "GOSemSim: An R package for measuring semantic similarity among GO terms and gene products," *Bioinformatics*, vol. 26, no. 7, pp. 976–978, 2010.
[88] G. Yu, L. G. Wang, Y. Han, and Q. Y. He, "ClusterProfiler: An R package for comparing biological themes among gene clusters," *OMICS J. Integrative Biol.*, vol. 16, no. 5, pp. 284–287, 2012.
[89] D. Warde-Farley, S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes, and A. Maitland, "The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function," *Nucleic Acids Res.* vol. 38, pp. W214–W220, Jul. 2010.
[90] R. A. Deyo, D. C. Cherkin, and M. A. Ciol, "Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases," *J. Clin. Epidemiol.*, vol. 45, no. 6, pp. 613–619, Jun. 1992, doi: 10.1016/0895-4356(92)90133-8.
[91] A. Elixhauser, C. Steiner, D. R. Harris, and R. M. Coffey, "Comorbidity measures for use with administrative data," *Med. Care*, vol. 36, pp. 8–27, Jan. 1998, doi: 10.1097/00005650-199801000-00004.
[92] C. A. Hidalgo, N. Blumm, A.-L. Barabási, and N. A. Christakis, "A dynamic network approach for the study of human phenotypes," *PLoS Comput. Biol.*, vol. 5, Apr. 2009, Art. no. e1000353, doi: 10.1371/journal.pcbi.1000353.
[93] M. A. Moni and P. Liò, "comoR: A software for disease comorbidity risk assessment," *J. Clin. Bioinform.*, vol. 4, p. 8, May 2014, doi: 10.1186/2043-9113-4-8.
[94] M. A. Moni and P. Liò, "How to build personalized multi-omics comorbidity profiles," *Frontiers Cell Develop. Biol.*, vol. 3, p. 28, Jun. 2015.
[95] F. Ronzano, A. Gutiérrez-Sacristán, and L. I. Furlong, "Comorbidity4j: A tool for interactive analysis of disease comorbidities over large patient datasets," *Bioinformatics*, vol. 35, no. 18, pp. 3530–3532, 2019.
[96] A. Gutiérrez-Sacristán, À. Bravo, A. Giannoula, M. A. Mayer, and F. Sanz, "comoRbidity: An R package for the systematic analysis of disease comorbidities," *Bioinformatics*, vol. 34, no. 18, pp. 3228–3230, 2018.
[97] M. A. Moni, H. Xu, and P. Lio, "CytoCom: A Cytoscape app to visualize, query and analyse disease comorbidity networks," *Bioinformatics*, vol. 31, pp. 969–971, Mar. 2014.
[98] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, and M. G. McDermott, "Enrichr: A comprehensive gene set enrichment analysis Web server 2016 update," *Nucleic Acids Res.*, vol. 44, no. W1, pp. W90–W97, 2016.
[99] A. Munshi and Y. R. Ahuja, "Genes associated with Alzheimer disease," *Neurol. Asia*, vol. 15, no. 2, pp. 109–118, 2010.
[100] M. R. Rahman, T. Islam, B. Turanli, T. Zaman, H. M. Faruquee, M. M. Rahman, M. N. Mollah, R. K. Nanda, K. Y. Arga, E. Gov, and M. A. Moni, "Networkbased approach to identify molecular signatures and therapeutic agents in Alzheimer's disease," *Comput. Biol. Chem.*, vol. 78, pp. 431–439, Feb. 2019.
[101] C. Van Cauwenberghe, C. Van Broeckhoven, and K. Sleegers, "The genetic landscape of Alzheimer disease: Clinical implications perspectives," *Genet. Med.*, vol. 18, no. 5, p. 421, 2016.
[102] M. Rahman, T. Islam, M. Shahjaman, T. Zaman, H. M. Faruquee, M. A. Jamal, F. Huq, J. M. Quinn, and M. A. Moni, "Discovering biomarkers and pathways shared by Alzheimer's disease and Ischemic stroke to identify novel therapeutic targets," *Medicina*, vol. 55, no. 5, p. 191, May 2019.
[103] S. Chen, P. Sayana, X. Zhang, and W. Le, "Genetics of amyotrophic lateral sclerosis: An update," *Mol. Neurodegener.*, vol. 8, no. 1, p. 28, 2013.
[104] H. K. Rana, M. R. Akhtar, M. B. Ahmed, P. Lio, J. M. Quinn, F. Huq, and M. A. Moni, "Genetic effects of welding fumes on the progression of neurodegenerative diseases," *Neurotoxicology*, vol. 71, pp. 93–101, Mar. 2019.
[105] M. R. Rahman, T. Islam, M. Shahjaman, J. M. Quinn, R. D. Holsinger, and M. A. Moni, "Identification of common molecular biomarker signatures in blood and brain of Alzheimers disease," *bioRxiv*, Jan. 2019, Art. no. 482828.
[106] C. Eykens and W. Robberecht, "The genetic basis of amyotrophic lateral sclerosis: Recent breakthroughs," *Adv. Genomics Genetics*, vol. 5, p. 327, Oct. 2015.
[107] P. V. S. D. Souza, W. B. V. D. R. Pinto, M. A. T. Chieia, and A. S. B. Oliveira, "Clinical and genetic basis of familial amyotrophic lateral sclerosis," *Arquivos Neuro-Psiquiatria*, vol. 73, no. 12, pp. 1026–1037, 2015.
[108] M. R. Rahman, T. Islam, F. Huq, J. M. Quinn, and M. A. Moni, "Identification of molecular signatures and pathways common to blood cells and brain tissue of amyotrophic lateral sclerosis patients," *Inform. Med. Unlocked*, vol. 16, p. 100193, Jan. 2019.
[109] M. C. Fahey, A. H. Maclennan, D. Kretzschmar, J. Gecz, and M. C. Kruer, "The genetic basis of cerebral palsy," *Develop. Med. Child Neurol.*, vol. 59, no. 5, pp. 462–469, 2017.
[110] L. Fagerberg, B. M. Hallström, P. Oksvold, C. Kampf, D. Djureinovic, J. Odeberg, M. Habuka, S. Tahmasebpoor, A. Danielsson, K. Edlund, and A. Asplund, "Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics," *Mol. Cellular Proteomics*, vol. 13, no. 2, pp. 397–406, 2014.
[111] W. F. Alsanie, V. Penna, M. Schachner, L. H. Thompson, and C. L. Parish, "Homophilic binding of the neural cell adhesion molecule CHL1 regulates development of ventral midbrain dopaminergic pathways," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 9368.
[112] GeneDx. *Genetic Testing for Epilepsy: A Guide for Patients*. Accessed: Oct. 26, 2019. [Online]. Available: https://www.genedx.com/wp-content/uploads/crm_docs/91040_Epilepsy-Patient-Guide.pdf
[113] C. T. Myers and H. C. Mefford, "Advancing epilepsy genetics in the genomic era," *Genome Med.*, vol. 7, no. 1, pp. 1–11, 2015.
[114] L. Arning and J. T. Epplen, "Genetic modifiers of Huntington's disease: Beyond CAG," *Future Neurol.*, vol. 7, no. 1, pp. 93–109, 2012.
[115] S. E. Baranzini, "Revealing the genetic basis of multiple sclerosis: Are we there yet?" *Current Opinion Genet. Develop.*, vol. 21, no. 3, pp. 317–324, 2011.
[116] C. G. Deluca, L. R. Yates, and A. D. Sadovnick, "Genetics and epidemiology of multiple sclerosis," *Premer Multiple Sclerosis*, Jan. 2016, doi: 10.1093/med/9780199341016.003.0002.
[117] T. Islam, M. R. Rahman, M. R. Karim, F. Huq, J. M. Quinn, and M. A. Moni, "Detection of multiple sclerosis using blood and brain cells transcript profiles: Insights from comprehensive bioinformatics approach," *Inform. Med. Unlocked*. vol. 27, Jun. 2019, Art. no. 100201.
[118] C. Schulte and T. Gasser, "Genetic basis of Parkinson's disease: Inheritance, penetrance, and expression," *The Application Clinical Genet.*, vol. 4, p. 67, 2011.
[119] M. K. Lin and M. J. Farrer, "Genetics and genomics of Parkinson's disease," *Genome Med.*, vol. 6, no. 6, p. 48, 2014.
[120] M. H. Ullman-Cullere and J. P. Mathew, "Emerging landscape of genomics in the electronic health record for personalized medicine," *Hum. Mutation*, vol. 32, pp. 512–516, 2011.

MD HABIBUR RAHMAN was born in Naogaon, Bangladesh, in 1984. He received the B.Sc. and M.Sc. degrees in computer science and engineering from Islamic University, Kushtia, Bangladesh. He is currently pursuing the Ph.D. degree with the Institute of Automation, Chinese Academy of Sciences, under the University of Chinese Academy of Sciences, Beijing, China. From 2013 to 2016, he was working as a Lecturer with the Department of Computer Science and Engineering, Islamic University. His research interests include bioinformatics, machine learning, artificial intelligence, and so on.

SILONG PENG received the B.S. degree in mathematics from Anhui University, in 1993, and the M.S. and Ph.D. degrees in mathematics from the Institute of Mathematics, Chinese Academy of Sciences (CAS), in 1995 and 1998, respectively. From 1998 to 2000, he worked as a Postdoctoral Researcher with the Institute of Automation, CAS. During this period, he was also a Visiting Scholar with the Department of Mechanics and Mathematics, Lomonosov Moscow State University, Russia. In 2000, he became a Full Professor of signal processing and pattern recognition with the Institute of Automation, CAS. His research interests include wavelets, multirate signal processing, and digital image processing.

XIYUAN HU received the B.S. and M.E. degrees in computer science from the Nanjing University of Science and Technology, Nanjing, China, in 2005 and 2007, respectively, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011. In July 2011, he became a member of the Institute of Automation, Chinese Academy of Sciences, where he is currently an Associate Professor. He was a Visiting Scholar with the ReyLab, Harvard Medical School, Boston, USA, in 2014. His research interests include adaptive signal processing, digital image processing, and compression.

CHEN CHEN received the B.Sc. degree in applied mathematics from the National University of Defense Technology, China, in 2005, and the M.Sc. and Ph.D. degrees in computer science from the University of Copenhagen, Denmark, in 2011 and 2013, respectively. She joined the Institute of Automation, Chinese Academy of Sciences, in 2015, where she is currently an Assistant Professor. She was a Visiting Scholar with AI Lab, Stanford University, in 2012. Her research interests include machine learning, pattern recognition, and medical image analysis.

SHAHADAT UDDIN received the bachelor's degree in computer science from the Bangladesh University of Engineering and Technology, the master's degree in information systems from Central Queensland University, Australia, and the Ph.D. degree in complex networks and health analytics from the University of Sydney, in 2011. He is currently a Senior Lecturer with the Complex Systems Research Group and the Project Management Program of the Faculty of Engineering, University of Sydney. He was a recipient of many research excellence awards, including the highly prestigious Dean's Research Award, University of Sydney, the Research Excellence Award, University of Sydney, the Director's Award, Central Queensland University, and the Excellence in Innovation Award, CRC Association, Brisbane.

JULIAN M. W. QUINN received the D.Phil. degree from the University of Oxford, U.K., in 1992. He was a Postdoctoral Researcher with the Bone, Joint and Cancer Group, St Vincent's Institute of Medical Research, Melbourne, Australia. He worked as a Senior Research Fellow with the Bone Biology Division, Garvan Institute of Medical Research, Darlinghurst, Australia. He has over 25 years of the postdoctoral experience running laboratory-based bone biology research, leading teams focused on determining drug and hormone effects on bone health. He has long experience in scientific and technical-scientific writing and data analysis.

MOHAMMAD ALI MONI received the Ph.D. degree in artificial intelligence and machine learning from the University of Cambridge, U.K., in 2014. From 2015 to 2017, he was a Postdoctoral Research Fellow with the Garvan Institute of Medical Research, Darlinghurst. He worked as an Associate Lecturer with the University of New South Wales, Australia. He was awarded the DVC (USyd) Fellowship with the University of Sydney, Australia, in 2017. His research interests include artificial intelligence, machine learning, data science, and clinical bioinformatics.